Title

Negative selection allows human primary fibroblasts to tolerate high somatic mutation loads induced by N-ethyl-N-nitrosourea

Authors

Ronald Cutler[1, 2, =], Johanna Heid[1, =], Shixiang Sun[1], Moonsook Lee[1], Alexander Y. Maslov[1, 3], Lei Zhang[4], Simone Sidoli[2], Xiao Dong[4, 5], Jan Vijg[1, *]

Affiliations

[1] Department of Genetics, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[2] Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461, USA

[3] Laboratory of Applied Genomic Technologies, Voronezh State University of Engineering Technology, Voronezh 394000, Russia

[4] Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN 55455, USA

[5] Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN 55455, USA

[=] Authors contributed equally to this work

[*] Correspondence to J.V. (jan.vijg@einsteinmed.edu)

Summary

With advanced sequencing methods, it has now been shown that thousands of mutations accumulate with age in most human tissues. While there is ample evidence that some mutations can clonally amplify and lead to disease, the total burden of mutations a cell can tolerate without functional decline remains unknown. Here we directly addressed this question by exposing serially cultured human primary fibroblasts to multiple, low doses of N-ethyl-N-nitrosourea and analyzing somatic mutation burden using single-cell whole genome sequencing. The results indicate that mitotically active cells can sustain at least ~56,000 single-nucleotide variants with only a slight adverse effect on growth rate. We provide evidence that such high levels of mutations are tolerated through negative selection against variants in gene coding and non-coding regions, and in sequences associated with genetic pathways for maintaining relevant cellular functions such as growth and cell identity. Since most tissues in adults are non-dividing, these results suggest that somatic mutations in the absence of negative selection may have functionally adverse effects.

Key words

Aging; Somatic mutations; Mutation burden; N-ethyl-N-nitrosourea; Single-cell whole-genome sequencing; Negative selection; Coding regions; Non-coding regions; Cell growth pathways; Human lung fibroblasts

42 **Introduction**

43 Aging is a complex process of continuous decline in cellular and tissue function that is
44 accompanied with increased disease risk[1]. DNA damage is thought to play a central
45 role in this process and may be a universal cause of aging[2]. Among the major molecular
46 consequences of DNA damage during aging are somatic DNA mutations. Somatic
47 mutations are the result of errors during repair or replication of a damaged DNA template
48 and can only be purged via the death of the cell or organism. The accumulation of somatic
49 mutations has since long been implicated as a major cause of aging[3, 4], which is
50 strongly supported by the age-related exponential increase of cancer, a disease known
51 to be caused by mutations[5]. Somatic mutations are difficult to analyze because of their
52 random nature, with very low variant allele frequencies. With the recent emergence of
53 single-cell and single-molecule assays, this situation has changed and we are now able
54 to quantitatively analyze most small mutations, including base substitutions and small
55 insertions and deletions, at high accuracy. Application of such methods has provided
56 ample evidence that somatic mutation rate is far higher than the germline mutation rate
57 and that thousands of somatic mutations accumulate with age in most, if not all, human
58 tissues[6-10], including postmitotic tissues such as the brain[11]. In addition, evidence
59 has emerged that postzygotic, somatic mutations are a cause of a wide variety of human
60 diseases other than cancer[12, 13]. Like cancer, many of these age-related effects may
61 be due to clonally amplified mutations[14-16]. Indeed, in blood, somatic mutations lead to

clonal hematopoiesis, which has been associated with increased risk of age-related diseases, such as cardiovascular disease, as well as overall mortality[17].

What thus far remains unknown is the possible functional impact, collectively, of thousands of random mutations accumulating in a non-cancerous cell, for example, by adversely affecting gene-coding or gene-regulatory sequences relevant for the given cell type. If so, one would expect a limit to the burden of random mutations a cell can tolerate or selection against mutations in important functional regions. Here, we directly tested this hypothesis by treating actively proliferating human primary fibroblasts with low, repeated doses of N-ethyl-N-nitrosourea (ENU), a powerful mutagen that induces almost exclusively single-nucleotide variants (SNVs). Our results show that 9 treatment cycles of ENU treatment over the course of ~24 population doublings linearly increased mutation frequency up to ~56,000 SNVs per cell, without a major effect on growth rate or cell death. We show that such a high tolerance for somatic mutations can be explained by negative selection against potentially deleterious mutations in gene coding and non-coding regions and in functional genomic regions associated with cell growth. These results suggest that accumulating somatic mutations may have adverse effects on cellular function when negative selection is not possible, as is the case in most cells and tissues in adult organisms, which are not mitotically active[18].

84 **Results**

85 Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell Growth

86 To determine the impact of increased somatic mutation burden on primary human cells
87 *in vitro*, we treated fetal lung fibroblasts (IMR-90) with a sublethal dose of 50 μg/mL of
88 ENU, which was followed by a recovery period of 7 days. For each cycle, 1 million cells
89 were passaged and treated again, and this process was repeated for 9 cycles (Figure 1A;
90 Methods). To monitor cell growth and toxicity, we measured cell number and apoptosis
91 rate at 3 and 7 days after each treatment to assess both short- and long-term effects of
92 the treatment (Methods). We measured cell growth by calculating population doublings
93 (PDs), and found that repeated treatment with ENU gradually decreased cell growth rate
94 over the course of the experiment, as at cycle 9 there were a total of 28.25±1.87 (mean
95 ± standard deviation) PDs for control cells compared to 24.1±1.34 PDs for ENU-treated
96 cells (Figure 1B). The effect of ENU treatment on cell growth was more pronounced at
97 the 3 day time point (Figure 1C) compared to the 7 day time point (Figure 1D), which
98 mostly explains the decrease in overall growth rate (Figure 1B). A larger effect on growth
99 rate at the 3 day timepoint was expected due to known adverse effects of the DNA
100 damage response on cell growth, immediately following mutagen treatment, which then
101 subsides over time. To identify the contribution of cell death to the slight decrease in
102 growth rate due to mutagen exposure, we measured markers of early or late apoptosis at
103 the 3 and 7 day time points (Figures 1E-F and 1G-H). The results showed a slight, but
104 consistent increase in early and late apoptosis in the ENU-treated group as compared to

105 the control cells at both time points. These results suggest that increased apoptosis may
106 contribute to the slight adverse effect on cell growth.

107 To analyze possible effects of a DNA damage response on the control and ENU-treated
108 cycling cells in more detail, we performed bulk mRNA-sequencing on samples collected
109 at cycles 1 and 9 for the control group and cycle 9 for the ENU-treated group (Figures
110 S1A-K; Table S1; Methods). As a positive control for the effect of cell passsaging, we first
111 compared the control cycle 9 and the control cycle 1 groups, which resulted in 979
112 significantly differentially expressed genes (DEGs) (Figure 2A; Table S2). We then
113 performed targeted gene set enrichment analysis (GSEA) which showed that cell cycle-
114 related genes were significantly decreased, while apoptosis-related genes were not
115 significantly changed (Figures 2B-C). As cells are known to become senescent during
116 passaging, we also tested for an enrichment of senescence-related genes, which were
117 found to be significantly increased, as expected (Figure 2D).

118 To assess the effects of the repeated ENU treatment, we compared the ENU-treated
119 cycle 9 and the control cycle 9 groups. Unexpectedly, this resulted in only 2 significant
120 DEGs (Figure 2C; Table S2; Supplementary Note 1). However, after performing GSEA
121 (which does not rely on p-value thresholds), we found that repeated ENU treatment had
122 a similar, albeit weaker effect as cell passaging, i.e., a significant decrease of cell cycle-
123 related genes and a significant increase in senescence-related genes (Figure 2F, 2H). In
124 contrast to the effect of cell passaging, we found a significant increase in the abundance

125 of apoptosis-related genes in the ENU-treated cells relative to the controls (Figure 2G),
126 confirming the results of the direct apoptosis measurements (Figures 1G-H).

127 Taken together, these phenotypic and molecular data suggest that the observed
128 decrease in cell growth rate due to repeated ENU treatment (Figure 1B) is, at least in
129 part, due to increased cell cycle inhibition (Figures 1C, 2F, 2H) and apoptosis (Figures
130 1E-F, 2G). As the resulting effect sizes due to repeated ENU treatment (i.e. mutation
131 accumulation) were relatively small, these results demonstrate the resilience of normal
132 primary cells proliferating in culture to this mutagen, which was previously known to have
133 low cell toxicity when used at low doses[19]. Next, we analyzed the effect of repeated
134 ENU treatment on somatic mutation burden.

135

136 Repeated Mutagen Treatment of Normal Cells Results in Extremely High Mutation
137 Burden

138 Somatic DNA mutations in normal (non-cancerous) cells, which are mostly unique to each
139 cell, cannot be distinguished from sequencing artifacts after bulk whole genome
140 sequencing. Hence, we used our established and highly accurate single-cell whole
141 genome sequencing assay [20, 21] to analyze mutation burden in 21 single-cells from
142 control and ENU-treated groups collected in 3 independent batches at cycles 3, 6, and 9
143 from the repeated ENU treatment experiment (Figure 1A, Table S3, Methods). Cells for
144 mutation analysis were collected at the 7-day time point after each treatment cycle, as
145 DNA damage responses should be abated by that time and all mutations fixed. After
146 correction for genome coverage, sensitivity, and technical artifacts (Figures S2A-G;

147 Methods), we found that the number of estimated single nucleotide variants (SNVs) per
148 cell in the control group started at 2,434±406 at cycle 1, which increased to 19,064±823
149 in the ENU-treated group at cycle 3, (~7.8-fold increase), and then 33,098±1,785 in the
150 ENU-treated group at cycle 6 (~1.7-fold increase compared to cycle 3). At cycle 9, SNVs
151 in the ENU-treated group reached a maximum of 55,954±4,066 SNVs (~1.7 fold increase
152 compared to the cycle 6 ENU-treated group and ~32 fold increase compared to cycle 1
153 controls). Importantly, there was no significant difference between control cells at cycle 1
154 and 9, ruling out that the increase in SNVs in ENU-treated cells was due to passaging.
155 We checked for the possibility of clonal expansion by looking at shared mutations
156 between cells within each group, but only detected 8 SNVs that were shared between 2
157 cells in the ENU-treated group at cycle 6 (Figure S2F). This indicates a shared common
158 ancestor, but not clonal expansion. The increase in SNVs in ENU-treated cells was
159 orthogonally validated using our recently developed single-molecule assay (Figures 3B
160 and S3A; Table S4; Supplementary Note 2)[22]. We conclude that our treatment
161 paradigm increased mutational load to high superphysiological levels without serious
162 toxicity, with the increase in SNVs over the course of the experiment occurring linearly in
163 a cycle-dependent manner (Figures 3C-D).

164 While ENU is known to mainly induce base substitution mutations through alkylation
165 damage, the estimated number of small insertions and deletions (INDELs) per cell was
166 also observed to increase due to ENU treatment, starting from 123±49 in the control group
167 at cycle 1, to 246±49 in the ENU-treated group at cycle 3 (~1.8-fold increase); however,
168 there was no further increase at cycles 6 and 9 (Figures S2I-J). As with SNVs, there was
169 no significant difference in the amount of estimated INDELs between control cells from

170 cycle 1 and cycle 9. INDELs have been previously observed to be induced by ENU, albeit
171 at low frequencies[23], and could be caused by DNA polymerase slippage during DNA
172 replication at sites of alkylation damage. While the lack of an increase in INDELs after
173 cycle 3 in the ENU-treated group may be suggestive of negative selection acting to
174 prevent a further increase, this finding remains inconclusive due to the low amount of
175 observed INDELs per cell (~20 INDELs per cell).

176 We then characterized the underlying SNV mutational processes in the aggregated
177 mutations from the cells at all cycles within the control and ENU-treated groups
178 (Supplementary Note 3; Methods). The ENU-treated group showed an increase in the
179 relative contribution of thymine-to-adenine (T>A) and thymine-to-cytosine (T>C) base
180 transversions in both the single-cell and single-molecule data (Figures 4A, S3B, S4A).
181 This is in keeping with previously published ENU-induced mutational spectra by us and
182 others[19, 22]. We then inferred SNV mutational signatures *de novo* using non-negative
183 matrix factorization with a reference set of mutations from previous single-cell whole
184 genome studies [20, 24-26] (Figures S4H-I) and identified two signatures, named SBSA
185 and SBSB, which were dominant in the cells of the ENU-treated and control groups,
186 respectively (Figure 4B). Importantly, we found that SBSA resembled the SIGNAL human
187 signature associated with ENU-treatment, validating the source of the induced mutations
188 (Figure 4C)[27]. We then compared our *de novo* signatures to those in the COSMIC
189 database and found that SBSB was most similar to SBS5 a 'clock-like' mutational
190 signature that has been previously found to correlate with cellular replicataion during
191 aging [24, 28] (Figure 4D). Though relatively few INDELs were observed (369 in total),

192 we also provide a characterization of their mutational spectra (Figure S4K;
193 Supplementary Note 3).

194

195 Protection Against ENU-induced Mutation Accumulation of Damaging Coding and Non-
196 Coding Variants

197 Based on our initial results showing that the mutation burden of primary cells could linearly
198 increase to an estimated ~56,000 SNVs per cell (Figure 3A), i.e., ~10 times the SNV
199 burden typically found in normal cells from aged tissues[6], with only a slight effect on cell
200 growth and death (Figures 1B, 1E-H), we hypothesized that the sustained cell survival
201 and growth could be explained in part by negative selection against harmful mutations.
202 Our dataset presented a unique opportunity to test this hypothesis as there were a total
203 of 108,600 SNVs observed, which could provide a sufficient amount of power for
204 statistical testing of selection pressure (Figure S2H). To do this, we first annotated the
205 SNVs for different types of coding region variants: synonymous (non-protein-altering) and
206 non-synonymous (protein-altering) using the Ensemble Variant Effect Predictor
207 program[29], and then determined the frequency of these variants (Methods). This
208 revealed significant cycle-dependent increases in both synonymous and non-
209 synonymous variants in the ENU-treated group relative to the controls, with non-
210 synonymous variants reaching up to 80±35.3 variants per cell (73%±5.6% of all coding
211 variants per cell) at cycle 9 in the ENU-treated cells (Figure 5A).

212 We then measured selection pressure by calculating Log_2(observed/expected)
213 (Log_2(O/E)) ratios of the variants observed within each group, which compares the

214 frequency of observed variants to the frequency of variants expected by chance, i.e.,
215 neutral selection. The number of expected variants was estimated for each cell by
216 generating 10,000 instances of randomly shuffled mutations, which controlled for variation
217 in genomic coverage, mutational signatures, and sequence context of the observed
218 mutations [30] (Methods). Since the main hindrance in testing for selection is the low
219 number of mutations in normal cells (i.e., the control group), we used simulations to
220 determine if the frequency of observed variants was high enough to obtain sufficient
221 power for statistical testing at various Log_2(O/E) ratios (Methods). The results of these
222 simulations indicated that many of the Log_2(O/E) ratios obtained in the ENU-treated
223 groups were sufficiently powered, while almost none of the Log_2(O/E) ratios in the control
224 groups were sufficiently powered (Figures S5A-B; Table S5). Thus, we focused our
225 analysis on the ENU-treated cells (see Figure S5 for analysis of control cells). Finally, to
226 account for the known effect of DNA repair to deplete mutations within transcribed regions
227 through transcription coupled repair (TCR) and mismatch repair (MMR)[31, 32], we
228 normalized the Log_2(O/E) ratios of variants within transcribed regions by the Log_2(O/E)
229 ratio of mutations within intronic regions (assumed to be independent of selection)
230 (Methods).

231 With this approach, we found that the Log_2(O/E) ratios of non-synonymous variants at
232 cycles 6 and 9 in the ENU-treated group were ~ -0.2, which was significantly lower than
233 what would be expected by chance and indicative of negative selection (Figures 5B and
234 S5C-E, right panels). These Log_2(O/E) ratios corresponded to a ~20% reduction in
235 mutation burden which was a decrease of 14.5±9.1, and 19.9±12.5 non-synonymous
236 variants per cell at cycles 6 and 9, respectively (Figure 5C). In regards to synonymous

237 variants, the frequency of these was too low, which rendered the Log_2(O/E) ratios of these
238 variants to be insufficiently powered (Figures 5B and S5C-E, left panels; Table S5).

239 We then sought to verify if negative selection against non-synonymous variants could
240 also be shown using a method to calculate dN/dS ratios (i.e. the dNdScv program), an
241 alternative measure of selection [33] (Methods). It should be noted that there are
242 important differences between our method to calculate Log_2(O/E) ratios and the dNdScv
243 method. Unlike our O/E method, which empirically estimates expected mutations via
244 genome-wide simulations accounting for coverage and mutational signatures, the
245 dNdScv method relies on probabilistic modeling, assuming uniform genomic coverage
246 and abundant mutations in coding regions. Therefore, the assumptions of this method
247 made it less suitable for our single-cell whole genome sequencing data type, i.e., variable
248 genome coverage and a limited number of coding mutations. As expected, in contrast to
249 Log_2(O/E) ratios, the dN/dS ratios were ~1 and indicative of neutral selection (Figure S6A;
250 See Supplementary Note 4 for detailed comparison between methods).

251 To identify the contribution of the various types of non-synonymous variants to the
252 negative selection signal observed in the ENU-treated group, we stratified the non-
253 synonymous variants into stop retained, missense, start lost, stop gained, and stop lost
254 variants and calculated their respective Log_2(O/E) ratios (Supplementary Note 6). This
255 revealed that missense variants made up the majority of non-synonymous variants in the
256 ENU-treated group (Figure 5D) and also that the negative selection signal was driven
257 almost exclusively by missense variants (Figures 5E and S5F-H).

258 To further test whether the apparent negative selection against non-synonymous variants
259 persisted after normalising for local mutation input and coverage, we examined the ratio

260 of non-synonymous to synonymous changes (N/S). Synonymous variants, assumed to
261 be functionally neutral, occur in the same sequence contexts and at the same depth as
262 non-synonymous variants, so they serve as an internal baseline for the true mutation rate
263 introduced by ENU. A deficit of non-synonymous variants relative to this neutral baseline
264 is therefore attributable to negative selection rather than to sequencing bias or differences
265 in mutagenesis. Although the raw N/S ratios were similar between control and treated
266 cells (Figure 5F), applying the O/E method to these ratios produced Log_2(O/E) ratios
267 significantly below zero at cycles 6 and 9 in the ENU-treated group (Figures 5G, S5I-J),
268 corroborating the ~20 % selective loss of non-synonymous variants (Figure 5B, right
269 panel).

270 In an attempt to gain insight in selection pressure at the gene level, we mapped the non-
271 synonymous variants to the COSMIC Cancer Gene Census gene set and identified a total
272 of 75 cancer driver genes containing nonsynonymous variants (Methods). However, low
273 variant counts of ~1 per gene on average essentially constrained meaningful conclusions
274 when assessing selection pressure with either the O/E or dNdScv methods (Table S6).

275 Next, we tested if variants in non-coding regions were under selection pressure. First, we
276 used the Variant Effect Predictor program [29] to annotate variants in non-coding regions,
277 i.e., introns, intergenic, 5'UTR, 3'UTR, and splice sites. As with the coding variants, non-
278 coding variants showed cycle-dependent increases in the ENU-treated group relative to
279 the controls, with variants in intronic and intergenic regions accounting for 49.9% and
280 35.1% of all non-coding variants, respectively (Figure 5H). For ENU-treated groups at all
281 cycles, Log_2(O/E) ratios of intronic region variants were ~ -0.08, significantly less than
282 what would be expected by chance (Figure 5I). Since the vast majority of introns (with the

283 exception of splice sites and other conserved non-coding elements) do not contain
284 functional sequences and are evolving at neutrality in the germline, but are still subject to
285 DNA repair such as TCR and MMR[31, 32], we interpret the negative Log_2(O/E) ratios in
286 these regions as a signal of DNA repair rather than negative selection. As described at
287 the beginning of this section, we use the intron Log_2(O/E) ratio to account for the
288 confounding effect of DNA repair on the Log_2(O/E) ratio. In intergenic regions for ENU-
289 treated groups at all cycles, we found that the Log_2(O/E) was ~0.13, significantly greater
290 than what would be expected by chance and indicative of mutation enrichment (Figure
291 5I). Similar to introns, the vast majority of intergenic regions lack functional sequences
292 and are evolving at neutrality in the germline. However, DNA repair activity in these
293 regions is known to be substantially reduced and likely explains the positive Log_2(O/E)
294 ratios[34, 35]. However, we cannot discount the possibility of an alternative interpretation
295 of these results by a possible limitation of the O/E analysis method (Supplementary Note
296 5).

297 Out of all the non-coding variants, only Log_2(O/E) ratios of variants occurring in 3'
298 untranslated regions (UTR), and splice regions in the ENU-treated group at cycle 9 were
299 significantly less than 0 (Figures 5I S5K-M), which was indicative of negative selection
300 and supported by previous studies on germline selection pressures[34, 36, 37]. The
301 Log_2(O/E) ratios of 3' UTRs and splice sites in the ENU-treated group at cycle 9
302 corresponded to ~25% reduction in mutation burden, a decrease of 19.3±12.5 and
303 6.2±2.7 variants per cell, respectively (Figure 5J).

304 Taken together, these results show that potentially damaging variants in both coding and
305 a subset of non-coding regions did not accumulate to levels that would be expected in the

306 context of the extremely high mutation burden induced by ENU treatment and instead
307 were subject to negative selection, resulting in protection of critical genes supporting
308 essential cellular functions *in vitro*. These results explain how the population of cells
309 exposed to repeated mutagen treatments was able to avoid a dramatic decline in growth
310 rate (Figure 1B), even as overall SNV burden continued to increase linearly throughout
311 the experiment (Figure 3C-D). Of note, the selection pressure analysis was limited to
312 mutations that were actually physically detected, which are ~20% of the ~56,000
313 estimated mutations at cycle 9 in the ENU-treated group (Figures 3A and S2H). Hence,
314 these results have relevance for the much smaller numbers of SNVs that are estimated
315 to accumulate during normal aging, e.g., from ~3,000 in human B lymphocytes to ~ 9,000
316 in human hepatocytes[6].

317

318 <u>Pathways Supporting Fibroblast Growth and Cell Identity are Protected from Mutation</u>
319 <u>Accumulation</u>

320 To gain insights on selective pressures at broader levels associated with cellular function,
321 we investigated if important gene sets that were active in proliferating human fibroblasts
322 were under negative selection. To do this, we analyzed the SNVs across annotated
323 functional genomic regions relevant to human lung fibroblasts (i.e. exons, upstream
324 promoters, and enhancers) that were grouped into various gene sets[38-40]. We then
325 calculated the frequency of SNVs within these regions as well as the Log_2(O/E) ratios
326 (while accounting for the effects of DNA repair by intron Log_2(O/E) ratio normalization, as
327 described above; Methods).

328 First, we analyzed non-expressed genes and expressed genes as determined by our bulk
329 mRNA-sequencing data (Figures S1A-B; Table S7; Methods). In keeping with the
330 previous analyses of other regions (e.g. exonic and intronic regions), we found cycle-
331 dependent increases of SNV frequency in non-expressed and expressed regions for the
332 ENU-treated group relative to the controls (Figure 6A). Log_2(O/E) ratios of SNVs in
333 expressed genes were significantly below 0 at all cycles in the ENU-treated group,
334 indicative of negative selection (Figures 6B and S7A-C, right panels). In contrast, the
335 Log_2(O/E) ratios of SNVs in non-expressed genes did not significantly deviate from the
336 expectation in the ENU-treated group at cycle 9, indicating neutral selection (Figures 6B
337 left panel, S6A-C, left panels). Next, we further subdivided the expressed genes into
338 nonessential and essential gene sets based on previous *in vitro* screens and repeated
339 the same analysis[41, 42]. This showed increased SNV frequency in the ENU-treated
340 group relative to the controls (Figure 6C), along with a Log_2(O/E) ratio significantly below
341 0 at cycle 9 in the ENU-treated group, indicative of negative selection (Figures 6D and
342 S7D-F, right panels). Of note, this is in agreement with global dN/dS ratios on truncating
343 mutations across 17 genes essential from a large-scale study sampling the normal oral
344 epithelium[43]. As negative selection is expected to be generally weak due to the diploid
345 nature of the human genome[33, 44-46], but stronger in haploinsufficient regions, we
346 subdivided expressed genes into haplosufficient (pLI > 0.1) and haploinsufficient (pLI <
347 0.9) gene sets based on a large scale study of protein-coding genetic variation[47]. As
348 expected, analysis of these gene sets showed increased SNV frequency in the ENU-
349 treated group relative to the controls (Figure 6E). Log_2(O/E) ratios were significantly below
350 0 in both haplosufficient and haploinsufficient gene sets for the ENU-treated cells at cycle

351 9. However, only the ENU-treated group at cycle 6 showed Log_2(O/E) ratios significantly
352 below 0 for the haploinsufficient gene set and not the haplosufficient gene set, suggesting
353 overall stronger negative selection of haploinsufficient genes, as would be expected
354 (Figures 6F and S7G-I). In further support of this, when the Log_2(O/E) was compared
355 between haplosufficient and haploinsufficient gene sets for ENU-treated cells at cycle 9,
356 we found that the Log_2(O/E) ratio was considerably lower for the haploinsufficient genes
357 (-0.37±0.16) as compared to the haplosufficient genes (-0.28±0.04). In summary, these
358 results suggest that damaging SNVs are selectively purged from functional genomic
359 regions based on expression status, essentiality, and dosage sensitivity, extending our
360 previous results on negative selection of damaging SNVs in coding and non-coding
361 regions.

362 Moving to the pathway level, we identified 399 gene ontology (GO) biological processes
363 as significantly enriched in the set of expressed genes from our bulk RNA-seq data
364 (Figure S7J; Table S8; Methods). As expected, the mean frequency of SNVs within each
365 pathway showed significant cycle-dependent increases of SNV frequency in the ENU-
366 treated group relative to the controls (Figure 6G; Table S9). At the global level, Log_2(O/E)
367 ratios of the identified pathways in the ENU-treated group showed a cycle-dependent
368 decrease, suggesting widespread negative selection at the pathway level (Figures 6H,
369 S7K-L; Table S9). At cycle 9 in the ENU-treated group, 6 pathways showed Log_2(O/E)
370 ratios significantly below 0, indicative of negative selection. The pathways under negative
371 selection were characterized as those related to basic cellular processes supporting lung
372 fibroblast cell growth *in vitro* and cellular identity, i.e., 'response to oxidative stress'
373 (GO:0006979), 'respiratory tube development' (GO:0030323), 'respiratory system

374 development' (GO:0006979), 'regulation of response to DNA damage stimulus'
375 (GO:2001020), 'proteasome-mediated ubiquitin-dependent protein catabolic process'
376 (GO:0043161), and 'cellular response to chemical stress' (GO:0062197) (Figures 6I-K;
377 Table S9). To ensure that the pathways we had identified to be under negative selection
378 were not false positives, we repeated the pathway analysis, but employed bootstrapping
379 to randomly shuffle genes amongst pathways before calculating the Log_2(O/E) ratios
380 (Figures S8A-C; Methods). The results of this served as a negative control, which had no
381 significant overlap with the pathways we identified to be under significant negative
382 slection pressure, providing a validation of our results (Figures S8D-E).

383 In summary, these results suggest that in the context of very high mutation burden,
384 pathways supporting oxidative stress, cellular identity, and protein homeostasis in
385 proliferating cells *in vitro*, are protected from the accumulation of SNVs by negative
386 selection. Importantly, these pathways are biologically significant, as they appear to be
387 relevant to the conditions of *in vitro* cell culture and to the unique charecteristics of lung
388 fibroblast, highlighting the functional importance of preventing mutation accumulation.
389 This shows how negative selection allows a population of dividing cells to survive
390 extremely high mutation burden and underscores the impact of random somatic SNVs on
391 essential cellular functions.

392

393 **Discussion**

394 Somatic mutations are known since the 1970s to be the main cause of cancer[48]. Mainly
395 because of the exponential increase in cancer risk with age[49], aging itself has been

396 suggested to be caused by the accumulation of somatic mutations[3, 8, 50]. Testing this
397 hypothesis has long been constrained by a lack of methods to quantitatively analyze
398 somatic mutations in normal tissues during aging. However, both single-cell and single-
399 molecule methods have recently been developed to accurately quantify randomly
400 accumulating SNVs and INDELs. Using these methods, we and others have
401 demonstrated that somatic mutation burden is far higher than expected based on the
402 known germline mutation frequency and increases with age up to several thousands of
403 mutations per cell in aged subjects depending on the tissue[6, 9]. However, a causal
404 relationship of increased mutation burden and age-related functional decline and disease
405 risk remains untested. Instead, attention has been focused on clonally amplified somatic
406 mutations and their causal relationship to non-cancerous and age-related diseases[14,
407 51].

408 In the present study, we tested for a possible limit to somatic mutation accumulation in a
409 cell culture model in which high levels of somatic mutations were induced by multiple low
410 doses of the powerful mutagen ENU (Figure 1A). In this system we were able to linearly
411 increase the somatic mutation burden to ~56,000 SNVs per cell, over 30-fold higher than
412 untreated control cells (Figure 3A). Remarkably, this supraphysiological mutation load
413 only slightly impacted cell growth rate (Figure 1B), suggesting that most cells can sustain
414 an astounding level of genetic damage. The most straightforward explanation for these
415 findings is that the mammalian genome is highly robust, and that even very high numbers
416 of point mutations are unlikely to adversely affect cellular function. Indeed, this would be
417 supported by results from a recent study in which as many as ~30,000 SNVs were

418 observed in intestinal crypts of elderly patients with genetic defects in DNA mismatch
419 repair[52].

420 An alternative explanation for the high tolerance of actively proliferating cells to mutation
421 accumulation is negative selection against mutations in genomic regions important for
422 cell survival. Evidence for negative selection against somatic mutations has been found
423 in tumors[46]. However, in cancer, positive selection is the dominant force, arising from
424 mutations in cancer driver genes[33]. Evidence for negative selection of somatic
425 mutations in normal cells is scarce because of the much lower frequency of somatic
426 mutations, which limits robust statistical testing. Our present analysis of untreated control
427 cells indeed suffered from this same limitation, and we were unable to convincingly
428 demonstrate negative selection of mutations in these cells (Figure S5D). However, our
429 ENU-treated cells did yield a sufficiently high number of SNVs to test for negative
430 selection (Figure S2B). With this, we found multiple lines of evidence for negative
431 selection acting on potentially deleterious coding and non-coding variants. While the
432 frequency of both synonymous and non-synonymous variants increased linearly with
433 ENU treatment (Figure 5A), deeper analysis using observed vs. expected ($Log_2(O/E)$)
434 ratios revealed a significantly lower frequency of high impact non-synonymous variants,
435 especially missense variants, in the ENU-treated cells as compared to our null model
436 (Figures 5A-E). This suggests that an actively proliferating population of fibroblasts has
437 the capacity to selectively purge the most harmful coding mutations from their genomes,
438 preserving cellular fitness despite the extremely high number of random mutations
439 accumulating genome wide, and agrees with the selection pressures found from
440 analyzing germline mutations[47, 53-55]. Additionaly, negative selection was also

441 detected in non-coding regions for variants occurring in 3' untranslated regions and splice
442 sites (which made up ~1.1% of all the observed non-coding variants), which is in
443 agreement with recent literature analyzing selective pressures in these regions [37]
444 (Figures 5I-J).

445 Our analysis of gene sets and pathways revealed negative selection to also be acting at
446 a broader level. In the primary lung fibroblasts treated with ENU, we found a significant
447 decrease in the frequency of mutations in genomic regions of actively expressed
448 pathways such as oxidative stress, respiratory system development, and response to
449 DNA damage (Figure 6F). As these pathways are relevant to our *in vitro* experimental
450 conditions (Figure 1A) and the cell type used in the experiments, this suggests that
451 despite the massive somatic mutational burden across the genome overall, somatic
452 mutations that occur in pathways critical for cellular survival have a damaging effect.

453 Our work provides the first evidence that very high numbers of somatic mutations,
454 collectively, may have a functional impact on normal cells independent of clonally
455 amplified mutations. However, because a supraphysioloigical amount of mutations was
456 required to detect negative selection, this puts into question whether there is an effect of
457 negative selection on the much lower numbers of somatic mutations that accumulate
458 during aging. In this respect, it should be noted that the actual number of SNVs measured
459 is much smaller that the ~56,000 estimated based on genome coverage and sensitivity,
460 i.e., no more than ~12,000 (Figure 3A and S2H). This is not that so much above the about
461 2,000 and 5,000 SNVs per cell, depending on the cell type analyzed, that have been
462 estimated to have accumulated at old age [6]. Hence, it is conceivable that negative
463 selection is a general mechanism to eliminate somatic mutations that adversely impact

464 cell function, similar to purifying selection in the germline. However, such a mechanism
465 would only work in actively proliferating cells when negative selection is possible, similar
466 to the situation in tumors. In normal adult subjects, cell division rate is very low with few
467 exceptions, most notably the lymphoid and intestinal systems [15]. In human B
468 lymphocytes, we previously demonstrated that SNVs accumulate with age much slower
469 in the functional part of the genome than in the genome overall, indicating negative
470 selection of mutations arising during the process of B cell aging[56]. By contrast, somatic
471 mutations accumulating in non-dividing adult cells will inevitably affect coding or
472 regulatory sequences associated with the particular function these cells are programmed
473 to carry out. However, such mutations cannot be cleared through negative selection in
474 the absence of a relevant cell fitness advantage (e.g., cell growth).

475 As a limitation of our study, we should point out that we only analyzed SNVs, and the few
476 small insertions and deletions induced by ENU. However, SNVs are much less likely to
477 have a functional impact than genome structural variants (SVs), including large deletions,
478 translocations, and copy number variants. Indeed, this is clearly indicated for INDELs,
479 which in our present study did not linearly accumulate during the ENU treatment (Figures
480 S2I-J). Thus far, there is very little information about somatic SVs, which cannot be
481 detected accurately, either using single-cell or single-molecule methods[57]. While their
482 relative frequency is much lower than that of SNVs, SVs can have a major phenotypic
483 impact, as indicated by germline SVs in human diseases[58]. SVs are much more likely
484 to adversely affect cell function than SNVs, but unless they result in cell death, a
485 mechanism to select against them in non-dividing cells *in vivo* during aging is difficult to
486 imagine.

Data Availability

Raw sequencing data has been deposited in the NCBI Sequence Read Archive under BioProject accession PRJNA1129131. Processed gene expression data has been deposited in the NCBI Gene Expression Omnibus database under accession GSE271867. Processed variant call files have been deposited in the NCBI dbSNP database and will be made available upon acceptance.

Code Availability

Scripts used for processing and analyzing data have been deposited at: figshare.com/articles/software/Negative_selection_allows_human_primary_fibroblasts_to_tolerate_high_somatic_mutation_loads_induced_by_N-ethyl-N-nitrosourea/26110453

Acknowledgements

This study was supported by National Institute of Health grants: T32GM007491 (R.C.) T32AG023475 (R.C.), U19AG056278 (J.V.), U01HL145560 (J.V.), U01ES029519 (J.V.), P01AG017242 (J.V.), P01AG047200 (J.V.), RF1AG068908 (J.V.), and P30AG038072 (J.V.). Other funding sources: US Department of Defense grant BC180689P1 (J.V.), Glenn Foundation for Medical Research Postdoctoral Fellowships in Aging Research (J.H.), The Michael J. Fox Foundation (J.V.), and The Paul F. Glenn Center for the Biology of Human Aging at the Albert Einstein College of Medicine (J.V.).

Author Contributions

A.Y.M., and J.V. conceived this study and designed the experiments. J.H., M.S., and A.Y.M performed the experiments. R.C., J.H., and S.S. analyzed the data. X.D. provided guidance on the data analysis. R.C., J.H., and J.V. wrote the manuscript.

Declaration of interests

M.L., A.Y.M., X.D., and J.V. are co-founders and shareholders of SingulOmics Corp. A.Y.M., and J.V. are co-founders of MutagenTech Inc. Other authors declare no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors employed ChatGPT for spelling and grammar checks during this work's preparation. They have reviewed and edited the content accordingly, assuming full responsibility for the publication's content.

STAR Methods

Cell culture and treatment

Human IMR-90 primary fibroblasts were obtained from the American Tissue Culture Collection (ATCC; LOT number 70020198). They were cultured in DMEM (Dulbecco's Modified Eagle Medium, from ATCC) supplemented with 10% fetal bovine serum (FBS,

528 Gibco/Life Technologies) and 1% Penicillin-Streptomycin solution (Gibco/Life
529 Technologies) at 3% oxygen and 37°C. Cells obtained from ATCC were at passage 11 at
530 arrival and were cultured for an additional 5 passages before the start of the experiment.
531 At the beginning of each cycle, $1*10^6$ human primary lung fibroblasts (IMR-90) cells were
532 plated and allowed to expand for 24 hours. Cells were then treated with 50 µg/mL N-ethyl-
533 N-nitrosourea (ENU) and allowed to recover for 3 days. Cells were then collected and
534 assayed for cell number viability for the 3-day time point. Of the collected cells, $1*10^6$ cells
535 were re-plated and then allowed another 4 days to recover and expand until they were
536 collected to assess cell number and viability for the 7-day time point. This procedure was
537 repeated for 9 consecutive cycles over the course of ~10 weeks. This experiment was
538 repeated 3 times to create 3 independent biological replicates. Cells were isolated at the
539 7-day time point at cycles 1, 3, 6, and 9 for somatic mutation analysis and at cycles 1 and
540 9 for transcriptome analysis.

541

542 <u>Cell growth and apoptosis assays</u>

543 Cells were counted using the ViaCount assay (Cytek) on a Guava EasyCyte flow
544 cytometer to assess cell number and viability at each time point. Apoptosis rates were
545 determined utilizing the Nexin assay (Cytek) on the same machine for experiments A and
546 B. Population doublings (PDs) is defined as how many times a population of cells as
547 doubled over a given period of time, as was calculated as: $PD = 3.32* (\log_{10} (\text{number of}$
548 $\text{cells counted}) - \log_{10} (\text{number of seeded cells}))$. The cumulative sum is then plotted. The
549 seeded cell number was always $1*10^6$ cells.

550

551 Bulk mRNA transcriptome analysis

552 RNA was extracted from frozen cell pellets using the Zymo DNA/RNA extraction kit
553 (Microprep plus D7005) according to the manufacturer's instructions and quantified using
554 a Qubit kit (Thermo Fisher Scientific, USA). Libraries for mRNA analysis were prepared
555 and sequenced using 150 paired-end mode on an Illumina Novaseq platform by
556 Novogene Corp Inc., CA. Raw sequence reads were processed using the nf-core RNA-
557 Seq pipeline (version 3.1.0) to obtain gene expression levels[59, 60]. Gene filtering was
558 performed using the zPKM program (Figure S1A)[61]. Count normalization and
559 differential expression testing was performed using DESeq2 [62] where the design was
560 set to test for the effect of the cycle-condition variable while controlling for the batch
561 variable (Supplementary Note 1). Gene set enrichment analysis (GSEA) was performed
562 using the clusterProfiler program [63] along with the SenMayo[64], KEGG DNA repair,
563 and KEGG apoptosis gene sets[65].

564

565 Single-cell isolation, whole-genome amplification, library preparation, and sequencing

566 Single-cells were isolated using the CellRaft AIR system (Cell Microsystems). The
567 CellRaft array was prepared following the manufacturer's instruction and washed three
568 times for 3 minutes with warm PBS before addition of cell suspension. 1,000 cells were
569 seeded in 1ml for one CellRaft array. Cells were allowed to settle for 3-4 hours at 37°C
570 before isolation. Then, the raft was carefully washed with warm medium to remove
571 potential debris and dead cells. Individual rafts containing one cell were deposited in PCR

572 tubes with 2.5 µl PBS and stored at -80°C until further processing. Single-cell whole-
573 genome amplification and library preparation for WGS was performed as previously
574 described [12]. In brief, single-cells were whole genome-amplified using SCMDA and
575 subject to quality control using a locus dropout test[66]. Libraries were prepared using a
576 NEBNext Ultra II FS kit (NEB). Library quantity and size were assessed with a Qubit kit
577 and TapeStation, respectively. WGS libraries were sequenced using the 150 paired-end
578 mode on an Illumina Novoseq platform by Novogene Corp Inc., CA.

579

580 Single-cell whole-genome read alignment and mutation calling

581 Data analysis was performed as previously described[21]. Briefly, raw sequencing data
582 underwent quality control using FastQC [67] and Trim Galore [68], followed by alignment
583 to the human hg19 reference genome using the GATK best practices pipeline [69]. This
584 included marking adapters, sequence alignment, base quality score recalibration, and
585 variant calling for germline variants using GATK HaplotypeCaller. SCcaller (version 2.0)
586 [21] was then employed to call somatic single-nucleotide variations (SNVs) and small
587 insertions and deletions (INDELs), adjusting for allelic amplification bias using known
588 germline variants (Figures S2A-C; Table S10). Estimated mutation burden was calculated
589 as follows:

590

591 $$\text{Estimated somatic SNVs|INDELs per cell} = \frac{\text{observed somatic SNVs|INDELs}}{\frac{surveyed\ genome\ (bp) * sensitivity}{size\ of\ genome\ (bp)}}$$

592

593 Mutation filtering and identification of technical artifacts in single-cell whole-genome
594 mutation data

595 Overlapping SNVs and INDELs were observed amongst the single-cells from batches 2
596 and 3 (Figures S2D-E; Table S3). Since this was not observed in the first batch, it is
597 unlikely that these mutations are the result of clonal amplification. Mutations that were
598 shared between all cells within a batch were thus considered to be germline and were
599 filtered out prior to downstream analysis (Figures S2F-G).

600 Technical artifacts inherent to single-cell MDA library preparation steps, such as
601 overheating during cell lysis or whole genome amplification, can be identified by their
602 specific mutational signatures scE and scF[70]. We checked for the presence of these
603 signatures and found scF to be present in the signature extracted from control cells at
604 cycle 1 (Figure S4A). The scF technical artifact is characterized by higher than expected
605 cytosine-to-thymine (C>T) transversions at GCN contexts (Figures S4C-D). Thus, control
606 cells from cycle 1 were excluded from the downstream mutational signature analyses.

607 In an attempt to account for the C>T transition artifacts in control cells at cycle 1, we
608 compared the proportion of the C>T transitions in control cells at cycle 1 to control cells
609 at cycle 9, which do not show any mutational signatures with similarity to the scF
610 sequencing artifact mutational signature [70]and have mutation spectra previously
611 observed for IMR90 cells [71]. We found that this artifact increased the proportion of C>T
612 transitions by 18%±1%. With this, we estimated the amount of artifactual C>T transitions
613 in control cells at cycle 1, subtracted this from the observed mutation burden, and then
614 re-estimated the SNVs per cell (Figure S4J). This better matches what would be expected
615 for cells being passaged in culture (i.e. insignificant difference) and is in line with

616 previously published data from our lab, considering the relatively low amount of cell
617 population doublings throughout the whole experiment (i.e. ~30) [71].

618

619 Mutation signature analysis

620 The MutationalPatterns program [72] use used to analyze mutational spectra and
621 signature analysis. We obtained a diverse set of samples from SomaMutDB [73] to use
622 as the background reference for the updated mutational signature analysis. These were
623 single-cell whole genome sequencing samples from 10 fibroblast cells [20], 199 brain
624 cells [26],and 53 lung cells [25], where genomic DNA was amplified with either Multiple
625 Displacement Amplification (MDA) or Primary Template-directed Amplification (PTA)
626 (Figure S4H).To determine the number of mutational signatures to be extracted from the
627 data, a non-negative matrix factorization (NMF) rank survey was employed which resulted
628 in 2 factors (Figures S4B, S4I). *De novo* signatures were extracted and then compared
629 to those from the COSMIC and SIGNAL databases using cosine similarity [74, 75]. We
630 also provide results of the NMF rank survey and *de novo* signature extraction when
631 excluding reference samples (Figures S4B-E).

632

633 Variant annotation

634 Observed mutations in each single-cell and simulations of single-cells (where mutations
635 were randomly re-distributed as described in the section below) were annotated using the
636 Ensembl Variant Effect Prediction program (version 110.1) and the given 'Consequences'
637 annotation was used for downstream analysis [29]. In the case where mutations were

638 predicted to lead to more than 1 unique consequence (i.e. multiple entries of the same
639 mutation), the highest ranking annotation that occurred in the resulting output was kept
640 which resulted in a 1:1 mapping of mutations and consequences. In the case where
641 mutations occurred in a region that overlapped multiple possible consequences (e.g.
642 splice region variant and intron variant), we classified this mutation to be consistent with
643 the consequence that occurred first in this annotation (e.g. splice region, if multiple
644 annotation was splice region variant and intron variant). For mutations annotated as splice
645 variants (splice_acceptor_variant, splice_donor_variant, splice_donor_5th_base_variant,
646 splice_region_variant,splice_donor_region_variant, splice_polypyrimidine_tract_variant),
647 we classified these mutations as 'splice.' Synonymous variants were classified as
648 mutations annotated as synonymous_variant and stop_retained_variant. Non-
649 synonymous variants were classified as mutations annotated as missense_variant,
650 start_lost, stop_gained, and stop_lost.

651

652 <u>Selection analysis of annotated variants</u>

653 Selection analysis was performed by first creating 10,000 simulations of each single-cell
654 in our dataset to be as the null expectation (i.e. neutral selection) where mutations are
655 randomly re-distributed throughout the genome. This was done using the
656 SigProfilerSimulator program[30], where for each single-cell, observed mutations were
657 randomly re-distributed across the genome while accounting for cell-specific sequencing
658 coverage, mutational signatures, and nucleotide contexts.

659 Log_2(observed/expected) ratios (Log_2(O/E)) were calculated for each single-cell, as well
660 as by group where mutations from all cells are pooled, by first summing the frequency of
661 annotated consequences of all single-cells and their respective simulations within each
662 group. The Log_2(O/E) ratio was then calculated with respect to each simulation (resulting
663 in 10,000 Log_2(O/E) ratios per cell/group) as follows:

664 $$Log_2(O/E) = Log_2 \frac{(observed\ consequence\ freq. + 1)}{(simulated\ consequence\ freq. + 1)}$$

665

666 A permutation test was then used to determine the p-value of the observed consequence
667 frequency where the simulated consequence frequency served as the null distribution
668 and is as follows:

669 $$p = \frac{1}{n} \sum_{i=1}^{n} I\ (simulated\ consequence\ freq._i \geq observed\ consequence\ freq.)$$

670 Where *p* is the permutation test p-value, *n* is the number of permutations (i.e. 10,000), *i*
671 is the *i*-th permutation, and *I* is the indicator function, which equals 1 if the condition inside
672 is true, and 0 otherwise.

673 We performed a power analysis to determine how many observed variants are required
674 to obtain reliable statistical results at varying Log_2(O/E) ratios. To do this, we used the
675 mutation data from ENU-treated cells to obtain the relationship between the standard
676 deviation in the number of expected variants given the number of observed variants for
677 common types of variants that were identified. We fit a gamma GLM with a log link to
678 model standard deviation as a function of log-transformed observed variant count. The
679 model showed excellent fit (residual deviance = 3.19 on 144 df vs. null deviance = 147.37

680 on 145 df; dispersion = 0.020), with log-transformed observed variant count as a highly
681 significant predictor irrespective of the variant type ($p < 2 \times 10^{-16}$) (Figure S5A). With this,
682 we then calculated Log_2(O/E) ratios by simulating 10,000 instances of expected variant
683 counts while varying the observed variant counts from 1-1,000. We then varied the
684 Log_2(O/E) ratios from 0-1 by multiplying the observed variant count by a scaling factor
685 before calculating the Log_2(O/E) ratio. Finally, we calculated p-values for the Log_2(O/E)
686 ratios using a permutation test as described previously.

687 The power analysis demonstrated how varying the observed variant count affects the p-
688 value of the Log_2(O/E) ratio at different levels. Importantly, this analysis provided us with
689 estimates of how many observed variants were required to achieve sufficient power (p-
690 value < 0.05) (i.e. obtain reliable statistical results) given a specific Log_2(O/E) (Figure
691 S5B, solid line). This threshold was then applied to the Log_2(O/E) results, at both the
692 levels of variant types and pathways, to determine which of these were amenable to
693 further statistical testing where the results of this can be seen in the figure legends (e.g.
694 Figure 5B; Table S5). For the Log_2(O/E) non-synonymous/synonymous calculation, we
695 determined if these were amenable to further statistical testing if either the non-
696 synonymous or synonymous Log_2(O/E) results in each group reached the established
697 observed variant threshold (Table S5).

698 When assessing for mutation selection, it is also necessary to account for the cell-intrinsic
699 bias in mutation frequency due to DNA repair mechanisms (i.e. transcription coupled
700 repair), which is known to reduce mutation frequency in transcribed regions and is
701 agnostic to damage that would result in synonymous or non-synonymous variants [76].
702 This bias in mutation frequency was indeed observed in our data and consistent with a

703 depletion of mutations due to DNA repair (Figure 5I). Thus, we normalized the Log_2(O/E)
704 ratio of coding variants (e.g. synonymous, missense, etc.) and non-coding variants (5'/3'
705 UTR and splice region) by the Log_2(O/E) ratio of mutations which occurred within intronic
706 regions and not in splice regions. We chose to use this normalization because the
707 occurrence of both exonic and intronic mutations is dependent on transcription coupled
708 repair [31] however, since intronic mutations do not alter the translated protein (with the
709 exception of splice site variants), they are not subject to selective mechanisms. In
710 addition, the majority of variants in our data occur within introns (~46.7%, Figure 5H),
711 allowing us to make more confident Log_2(O/E) ratio for these mutations as compared to
712 using the Log_2(O/E) ratio of synonymous mutations which only made up ~0.23% of all
713 variants. A comparison of unnormalized and normalized Log_2(O/E) ratio for non-
714 synonymous and synonymous variants is shown for example by comparing Figures 5B
715 to S5C, where the effect of normalization can be seen to slightly increase the Log_2(O/E)
716 ratio.

717 To estimate the number of variants that are protected against resulting from negative
718 selection, we subtracted the average number of variants from simulated samples from
719 the number of variants in the respective observed sample.

720

721 <u>Selection analysis of gene sets and pathways</u>

722 Non-expressed and expressed genes were determined using bulk mRNA transcriptome
723 data from cycle 9 control and ENU-treated groups using the zPKM program (Figure S2A;
724 Table S7) [61]. Nonessential and essential control gene sets were obtained from the

725 DepMap database (depmap.org) [77], where the nonessential gene set was the negative
726 controls from Hart *et al* (2015) and the essential gene set was the intersection of the the
727 essential gene sets resulting from the Hart *et al* (2015) and Blomen *et al* (2014) studies
728 [42, 78]. These gene sets were filtered for expressed genes by overlapping them with the
729 list of expressed genes we determined previously (Table S7). Haplosufficient and
730 haploinsufficient gene sets were obtained from the ExAC database where haplosufficient
731 genes were defined as pLI < 0.1 and haploinsufficient genes were defined as pLI > 0.9
732 [47]. These gene sets were also filtered for expressed genes by overlapping them with
733 the list of expressed genes we determined previously (Table S7). Pathway genes were
734 obtained by performing over representation analysis for gene ontology biological
735 processes using the clusterProfiler program [63] on the list of expressed genes that
736 determined previously (Table S8). For all gene sets, genomic coordinates of exons,
737 introns, promoters (5KB upstream of transcription start site), and annotated enhancers
738 [38] were retrieved using the GenomicRanges (version 1.60.0) software [79].

739 For each gene set, two Log_2(O/E) ratios were calculated. One for the version of the gene
740 set where the intronic regions were included, and the other for the version of the gene set
741 intronic regions were excluded. To account for DNA repair (as previously described for
742 Log_2(O/E) calculation of coding and non-coding variants), the Log_2(O/E) ratios of each
743 gene set was normalized by subtracting Log_2(O/E) ratio of the version that included
744 introns from the version that excluded introns. A permutation test was then used to
745 estimate p-values, as described above, and were adjusted for multiple hypothesis testing
746 using the Benjamini-Hochberg method.

747 To serve as a negative control for the pathway analysis, for each enriched biological
748 process gene set, we randomly shuffled the genes between all other identified biological
749 processes gene sets and created 10,000 randomly permuted gene sets. Average
750 Log_2(O/E) ratios were then calculated for the 10,000 permutations of each gene set
751 followed by p-value estimation, as described above.

752

753 Single-molecule mutation analysis

754 DNA was extracted from frozen cell pellets using the Zymo DNA/RNA extraction kit
755 (Microprep plus D7005) according to the manufacturer's instructions and quantified using
756 a Qubit kit (Thermo Fisher Scientific, USA). Library preparation for single-molecule
757 mutation sequencing (SMM-Seq) was performed as previously described[22]. In brief,
758 200 ng gDNA was used as input for library construction and then fragmented using
759 restriction endonucleases Alu I and Mlu CI, followed by purification with AMPure XP
760 beads. The fragmented DNA was then processed with the NEBNext Ultra II DNA Library
761 Prep Kit for Illumina, utilizing a custom adapter. Post size selection with AMPure XP
762 beads, the product was quantified with a Qubit kit and analyzed on a TapeStation
763 (Agilent). For library preparation, samples were diluted based on molar concentration,
764 and 1 µl of the diluted sample was used as a template in a pulse-RCA reaction with a
765 strand displacement polymerase. The reaction product was purified, PCR-amplified, and
766 further purified before quantification and pooling for sequencing. Conventional whole-
767 genome sequencing (WGS) libraries that are used to identify germline variants were
768 similarly prepared by PCR amplification of adapter-ligated samples, followed by
769 purification, quantification, pooling, and sequencing as described above. WGS libraries

770 were sequenced using 150 paired-end mode on an Illumina Novoseq platform by
771 Novogene Corp Inc., CA.

772

773 Single-molecule whole-genome read alignment and mutation calling

774 Data analysis for SMM-Seq data was performed as previously described[22]. In brief, raw
775 sequence reads were adapter- and quality-trimmed, then aligned to the hg19 reference
776 genome. Realignment and recalibration were performed based on known indels. For
777 variant calling, filters were applied to ensure only high-quality reads (mapping quality ≥
778 60, in proper pairs, no secondary alignments) were considered. Positions covered by at
779 least seven reads from each strand within a UMI family and covered at least 20× in regular
780 sequencing were deemed qualified. Variants were identified if all reads within a UMI
781 family showed the same base, differing from the reference genome. Germline variants
782 were filtered out using a list of SNPs from the sample and dbSNP [80]Variants were
783 rejected if detected in different UMI families or conventional WGS data containing
784 germline variants. Somatic SNV frequency was calculated as the ratio of identified
785 variants to the total number of qualified positions. See Supplementary Note 2 for details
786 of analysis.

787

788 **Supplementary Notes**

789 Supplementary Note 1

790 Strong batch effects were found to be present during athe analysis of the bulk mRNA-
791 Seq data (Figures S1G-I), which could mask the effects of interest when comparing the
792 different groups. While the batch effect was corrected for during the differential expression
793 (DE) analysis by accounting for the batch variable in the generalized linear model utilized
794 by DESeq2 (Figures S1J-K), we also performed the DE analysis of ENU Cycle 9 vs
795 Control Cycle 9 after removing samples from the Control Cycle 1 group. After batch
796 correction, separation between the groups could be seen from principal component
797 analysis (Figure S1L). However, a very small number of genes were differentially
798 expressed, validating our findings then including the Control Cycle 3 group (Figure S1M).

799

800 Supplementary Note 2

801 We used our recently developed single-molecule assay (SMM-seq) to verify mutation
802 burden in 14 samples from control and ENU-treated groups collected in 3 independent
803 batches at cycles 1, 3, 6, and 9 at the 7-day time point (Figure 1A, Table S4). The number
804 of estimated SNVs per mega base (MB) increased from $4.03*10^{-1}\pm1.23*10^{-3}$ per MB in
805 the control group at cycle 1 to $1.95\pm9.55*10^{-2}$ in the ENU-treated group a cycle 3, i.e. a
806 4.8-fold increase (Figure 3B). At cycle 9, there was a ~14-fold increase in estimated SNVs
807 per MB in the ENU-treated group relative to the controls, where the total amount of
808 estimated SNVs per cell in the ENU-treated group reached $5.58\pm5.27*10^{-1}$ SNVs per MB.
809 Importantly, as found with the single-cell assay, we could not detect a significant
810 difference between controls from cycle 1 and cycle 9, which had an average frequency of
811 $4*10^{-1}\pm1*10^{-3}$and $5*10^{-1}\pm4*10^{-2}$ SNVs per MB, respectively. This indicates that the
812 culturing and passaging of the cells itself had very little impact on the mutational load.

813 Like for the single-cell assay, the accumulation of SNVs between the 1st and 9th cycle was
814 best fit by a linear model without evidence that a plateau had been reached (Figure 3D).
815 These results were highly correlative with the single-cell whole-genome sequencing
816 assay (Figure S3A), although the estimates tended to be lower. As the SMM-Seq method
817 uses a totally different single-molecule approach to overcome the issue of distinguishing
818 somatic mutations from sequencing errors and thus is subject to different technical
819 biases, it is expected that the estimated somatic mutation burden will differ between
820 SMM-Seq and SCMDA results, although the relative trends should be the same.
821 Moreover, the parameters which are used for the SMM-Seq analysis are tuned to
822 enhance specificity at the expense of sensitivity, which explains why SMM-Seq results
823 agree with SCMDA results at mutation burdens which are expected to occur
824 physiologically (< 1 SNV per 1 Mbp) and become conservative at higher burdens (> 1
825 SNV per 1 Mbp). Mutation spectra analysis indicated an increase in A>T transitions as
826 expected from ENU treatment (Figure S3B). Of note, the SMM-Seq obtains far fewer
827 observed mutations as compared to SCMDA, which constrains measuring selection
828 pressure (Table S4). Hence, while the SMM-Seq method essentially confirmed the single-
829 cell results, we based all our downstream analyses solely on the single-cell data.

830

831 Supplementary Note 3

832 Though relatively few INDELs were observed (369 total), the mutational spectra in ENU-
833 treated cells were characterized by an increase in thymine INDELs as well as INDELs
834 with a 5+bp deletion and microhomology relative to the controls (Figure S4E). Moreover,
835 a single *de novo* INDEL mutational signature could be extracted, which weakly resembled

836 COSMIC ID1 and ID2 (cosine similarity 0.59 & 0.54, respectively). Both have been
837 suggested to be due to slippage of the template strand during DNA replication in addition
838 to having “clock-like” behavior.

839

840 Supplementary Note 4

841 To test if the negative selection finding could be recapitulated with other software
842 calculates selection pressure, we used the dNdScv software with the default parameters
843 and as input we used the mutations from the control and ENU-treated groups. This
844 resulted in a global dN/dS ratio of 1 for all mutation types in all the ENU-treated groups,
845 indicative of neutral selection (Figure S6A). As these results conflicted with our own
846 results using the O/E method, we sought to determine what the differences were between
847 the two methods and if either of these was better suited to analyzing single-cell whole
848 genome sequencing data. We compared the counts of observed and expected for
849 synonymous (Figures S6B-F) and missense variants (Figures S6G-K) between the
850 methods, which led us to identify the main discrepancy: the approaches used to estimate
851 the expected number of variants (as hypothesized by the reviewer). The dndscv approach
852 calculates expected variants per gene based on probabilistic models of mutation rates
853 and sequence contexts[33]. In contrast, the O/E method uses SigProfilerSimulator to
854 empirically estimate expected variants by simulating mutations randomly across the
855 genome while accounting for coverage, sequence context, and mutational signatures[30].
856 Crucially, the dndscv model makes assumptions about uniform genomic coverage and
857 high mutation counts, conditions violated in our SCWGS data as it is characterized by
858 uneven coverage and relatively few coding mutations per cell (Figure S6A). Finally, an

859 argument against the null model generated by SigProfilerSimulator in the O/E method is
860 that due to its simplicity, it might not accurately reflect neutral selection as one would
861 expect (i.e. the null model does not represent neutral selection). However, when using
862 the 1,000 instances of simulated mutations generated from SigProfilerSimulator (based
863 on observed mutations from the control and ENU groups, respectively) as input to the
864 dndscv program, this produced a global dN/dS ratio of 1 for all variants, as would be
865 expected (Figure S6L). This means that the null model generated by SigProfilerSimulator
866 accurately models neutral selection pressure. In summary, while dN/dS is a powerful
867 metric for highly mutated cancer genomes, it lacks sensitivity to detect negative selection
868 in scenarios like ours with limited coding mutations and variable coverage across the
869 genome which is inherent to SCWGS. We conclude that empirically estimating expected
870 variants via simulations, as done in our O/E analysis, is better suited for such situations.

871

872 Supplementary Note 5

873 Given that there are enough observed variants for sufficient statistical power and after
874 accounting for DNA repair (Methods), negative Log_2O/E ratios suggest negative
875 selection because the frequency of observed variants is less than what would be
876 expected according to the null model. In contrast, positive Log_2O/E ratios do not
877 necessarily positive selection. This is because in the calculation of the Log_2O/E ratio,
878 while mutations are randomly re-distributed in the simulated samples according to the
879 null model, the total number of variants remains the same between observed and
880 simulated samples. This creates the potential for a see-saw effect as a depletion of

881 mutations in a large regions (i.e., introns, which make up ~25% of the genome) must be
882 balanced out. The consequence of this may be an inflated estimate of the Log_2O/E ratio
883 for other large regions (i.e., intergenic regions, which make up ~50% of the genome)
884 (Figure 5I), which warrants caution for the interpretation of these results. Despite this
885 potential limitation of the O/E method, we do believe that the resulting positive Log_2O/E
886 ratios for intergenic regions is reflective of the underlying biology, as this is supported by
887 previous evidence that intergenic regions lack DNA repair activity and are under neutral
888 selection[34, 35, 81].

889

890 Supplementary Note 6

891 Out of the 37 detected stop gain variants, 4 of these were in 4 different cancer driver
892 genes and had dN/dS ratios from ~70-1,000 when analyzing selection using dndscv,
893 though they were not significant after multiple hypothesis correction. However, these
894 particular variants could possibly explain the positively trending $Log_2(O/E)$ ratio seen for
895 stop gain variants in the ENU-treated cells (Figure 5D). These variants were found in
896 *ABL1* and *CDC73*, which are tumor suppressors while *CREB3L2* and *COL1A1* are
897 oncogenes. Thus, it may be possible that stop gain variants in ABL1 and CDC73 may
898 have provided a slight growth advantage that led it to being picked up in our analysis.
899 However, repeated passaging during ENU treatment creates a strong bottleneck that
900 limits clonal expansion. While 8 shared SNVs were detected between 2 cells, the
901 presence of ~9,000 unique SNVs per cell indicates these arose independently, arguing
902 against large clonal expansion (Supplementary Figure 2F). It will be interesting for further

903 studies which analyze larger numbers of somatic mutations in normal cells to see if this

904 may be true.

905

906 **References**

907 1. López-Otín, C., et al., *Hallmarks of aging: An expanding universe.* Cell, 2023.
908 **186**(2): p. 243-278.
909 2. Schumacher, B., et al., *The central role of DNA damage in the ageing process.*
910 Nature, 2021. **592**(7856): p. 695-703.
911 3. Failla, G., *The aging process and cancerogenesis.* Ann N Y Acad Sci, 1958. **71**(6):
912 p. 1124-40.
913 4. Szilard, L., *ON THE NATURE OF THE AGING PROCESS.* Proc Natl Acad Sci U
914 S A, 1959. **45**(1): p. 30-45.
915 5. Laconi, E., F. Marongiu, and J. DeGregori, *Cancer as a disease of old age:*
916 *changing mutational and microenvironmental landscapes.* Br J Cancer, 2020.
917 **122**(7): p. 943-952.
918 6. Ren, P., X. Dong, and J. Vijg, *Age-related somatic mutation burden in human*
919 *tissues.* Front Aging, 2022. **3**: p. 1018119.
920 7. Franco, I. and M. Eriksson, *Reverting to old theories of ageing with new evidence*
921 *for the role of somatic mutations.* Nat Rev Genet, 2022. **23**(11): p. 645-646.
922 8. Franco, I., G. Revêchon, and M. Eriksson, *Challenges of proving a causal role of*
923 *somatic mutations in the aging process.* Aging Cell, 2022. **21**(5): p. e13613.
924 9. Franco, I., et al., *Whole genome DNA sequencing provides an atlas of somatic*
925 *mutagenesis in healthy human cells and identifies a tumor-prone cell type.*
926 Genome Biol, 2019. **20**(1): p. 285.
927 10. Franco, I., et al., *Somatic mutagenesis in satellite cells associates with human*
928 *skeletal muscle aging.* Nat Commun, 2018. **9**(1): p. 800.
929 11. Lodato, M.A., et al., *Aging and neurodegeneration are associated with increased*
930 *mutations in single human neurons.* Science, 2018. **359**(6375): p. 555-559.
931 12. Erickson, R.P., *Somatic gene mutation and human disease other than cancer: an*
932 *update.* Mutat Res, 2010. **705**(2): p. 96-106.
933 13. Li, C. and S.M. Williams, *Human Somatic Variation: It's Not Just for Cancer*
934 *Anymore.* Current Genetic Medicine Reports, 2013. **1**(4): p. 212-218.
935 14. Abyzov, A., et al., *One thousand somatic SNVs per skin fibroblast cell set baseline*
936 *of mosaic mutational load with patterns that suggest proliferative origin.* Genome
937 Res, 2017. **27**(4): p. 512-523.
938 15. Laurie, C.C., et al., *Detectable clonal mosaicism from birth to old age and its*
939 *relationship to cancer.* Nat Genet, 2012. **44**(6): p. 642-50.
940 16. Abyzov, A., et al., *Somatic copy number mosaicism in human skin revealed by*
941 *induced pluripotent stem cells.* Nature, 2012. **492**(7429): p. 438-42.
942 17. Jaiswal, S., et al., *Age-related clonal hematopoiesis associated with adverse*
943 *outcomes.* N Engl J Med, 2014. **371**(26): p. 2488-98.
944 18. Sender, R. and R. Milo, *The distribution of cellular turnover in the human body.* Nat
945 Med, 2021. **27**(1): p. 45-48.

946 19. Barbaric, I., et al., *Spectrum of ENU-induced mutations in phenotype-driven and*
947 *gene-driven screens in the mouse.* Environ Mol Mutagen, 2007. **48**(2): p. 124-42.
948 20. Dong, X., et al., *Accurate identification of single-nucleotide variants in whole-*
949 *genome-amplified single cells.* Nat Methods, 2017. **14**(5): p. 491-493.
950 21. Zhang, L., et al., *Analyzing somatic mutations by single-cell whole-genome*
951 *sequencing.* Nat Protoc, 2024. **19**(2): p. 487-516.
952 22. Maslov, A.Y., et al., *Single-molecule, quantitative detection of low-abundance*
953 *somatic mutations by high-throughput sequencing.* Sci Adv, 2022. **8**(14): p.
954 eabm3259.
955 23. Watson, D.E., M.L. Cunningham, and K.R. Tindall, *Spontaneous and ENU-*
956 *induced mutation spectra at the cII locus in Big Blue Rat2 embryonic fibroblasts.*
957 Mutagenesis, 1998. **13**(5): p. 487-97.
958 24. Alexandrov, L.B., et al., *The repertoire of mutational signatures in human cancer.*
959 Nature, 2020. **578**(7793): p. 94-101.
960 25. Huang, Z., et al., *Single-cell analysis of somatic mutations in human bronchial*
961 *epithelial cells in relation to aging and smoking.* Nat Genet, 2022. **54**(4): p. 492-
962 498.
963 26. Miller, M.B., et al., *Somatic genomic changes in single Alzheimer's disease*
964 *neurons.* Nature, 2022. **604**(7907): p. 714-722.
965 27. Kucab, J.E., et al., *A Compendium of Mutational Signatures of Environmental*
966 *Agents.* Cell, 2019. **177**(4): p. 821-836 e16.
967 28. Hwang, T., et al., *Comprehensive whole-genome sequencing reveals origins of*
968 *mutational signatures associated with aging, mismatch repair deficiency and*
969 *temozolomide chemotherapy.* Nucleic Acids Res, 2025. **53**(1).
970 29. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* Genome Biol, 2016.
971 **17**(1): p. 122.
972 30. Bergstrom, E.N., et al., *Generating realistic null hypothesis of cancer mutational*
973 *landscapes using SigProfilerSimulator.* BMC Bioinformatics, 2020. **21**(1): p. 438.
974 31. Heilbrun, E.E., M. Merav, and S. Adar, *Exons and introns exhibit transcriptional*
975 *strand asymmetry of dinucleotide distribution, damage formation and DNA repair.*
976 NAR Genom Bioinform, 2021. **3**(1): p. lqab020.
977 32. Frigola, J., et al., *Reduced mutation rate in exons due to differential mismatch*
978 *repair.* Nat Genet, 2017. **49**(12): p. 1684-1692.
979 33. Martincorena, I., et al., *Universal Patterns of Selection in Cancer and Somatic*
980 *Tissues.* Cell, 2017. **171**(5): p. 1029-1041.e21.
981 34. Dukler, N., et al., *Extreme purifying selection against point mutations in the human*
982 *genome.* Nat Commun, 2022. **13**(1): p. 4312.
983 35. Yurchenko, A.A., et al., *Genomic mutation landscape of skin cancers from DNA*
984 *repair-deficient xeroderma pigmentosum patients.* Nat Commun, 2023. **14**(1): p.
985 2561.
986 36. Denisov, S.V., et al., *Weak negative and positive selection and the drift load at*
987 *splice sites.* Genome Biol Evol, 2014. **6**(6): p. 1437-47.
988 37. Findlay, S.D., L. Romo, and C.B. Burge, *Quantifying negative selection in human*
989 *3' UTRs uncovers constrained targets of RNA-binding proteins.* Nat Commun,
990 2024. **15**(1): p. 85.

991 38. Boix, C.A., et al., *Regulatory genomic circuitry of human disease loci by integrative*
992 *epigenomics.* Nature, 2021. **590**(7845): p. 300-307.
993 39. Carlson M, M.B. *TxDb.Hsapiens.UCSC.hg19.knownGene: Annotation package for*
994 *TxDb object(s).* 2015; Available from:
995 https://bioconductor.org/packages/release/data/annotation/html/TxDb.Hsapiens.
996 UCSC.hg19.knownGene.html.
997 40. Boix, C.A. *EpiMap Repository*. 2023; Available from:
998 https://compbio.mit.edu/epimap/.
999 41. Wang, T., et al., *Identification and characterization of essential genes in the human*
1000 *genome.* Science, 2015. **350**(6264): p. 1096-101.
1001 42. Hart, T., et al., *High-Resolution CRISPR Screens Reveal Fitness Genes and*
1002 *Genotype-Specific Cancer Liabilities.* Cell, 2015. **163**(6): p. 1515-26.
1003 43. Lawson, A.R.J., et al., *Somatic mutation and selection at epidemiological scale.*
1004 medRxiv, 2024: p. 2024.10.30.24316422.
1005 44. Vorontsov, I.E., et al., *Negative selection maintains transcription factor binding*
1006 *motifs in human cancer.* BMC Genomics, 2016. **17 Suppl 2**(Suppl 2): p. 395.
1007 45. Bányai, L., et al., *Use of signals of positive and negative selection to distinguish*
1008 *cancer genes and passenger genes.* Elife, 2021. **10**.
1009 46. Zapata, L., et al., *Negative selection in tumor genome evolution acts on essential*
1010 *cellular functions and the immunopeptidome.* Genome Biol, 2018. **19**(1): p. 67.
1011 47. Lek, M., et al., *Analysis of protein-coding genetic variation in 60,706 humans.*
1012 Nature, 2016. **536**(7616): p. 285-91.
1013 48. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976.
1014 **194**(4260): p. 23-8.
1015 49. Group, U.S.C.S.W., *US cancer statistics: 1999–2009 incidence and mortality web-*
1016 *based report.* 2013, USDHHS, CDC: Atlanta GA.
1017 50. Burnet, F.M., *Intrinsic mutagenesis: a genetic basis of ageing.* Pathology, 1974.
1018 **6**(1): p. 1-11.
1019 51. Ng, S.W.K., et al., *Convergent somatic mutations in metabolism genes in chronic*
1020 *liver disease.* Nature, 2021. **598**(7881): p. 473-478.
1021 52. Robinson, P.S., et al., *Increased somatic mutation burdens in normal human cells*
1022 *due to defective DNA polymerases.* Nat Genet, 2021. **53**(10): p. 1434-1442.
1023 53. Karczewski, K.J., et al., *The mutational constraint spectrum quantified from*
1024 *variation in 141,456 humans.* Nature, 2020. **581**(7809): p. 434-443.
1025 54. Zhang, X., et al., *Genetic constraint at single amino acid resolution in protein*
1026 *domains improves missense variant prioritisation and gene discovery.* Genome
1027 Med, 2024. **16**(1): p. 88.
1028 55. Gudkov, M., L. Thibaut, and E. Giannoulatou, *Context-adjusted proportion of*
1029 *singletons (CAPS): a novel metric for assessing negative selection in the human*
1030 *genome.* NAR Genom Bioinform, 2024. **6**(3): p. lqae111.
1031 56. Zhang, L., et al., *Single-cell whole-genome sequencing reveals the functional*
1032 *landscape of somatic mutations in B lymphocytes across the human lifespan.* Proc
1033 Natl Acad Sci U S A, 2019. **116**(18): p. 9014-9019.
1034 57. Hård, J., et al., *Long-read whole-genome analysis of human single cells.* Nature
1035 Communications, 2023. **14**(1): p. 5164.

1036 58. Yang, L., *A Practical Guide for Structural Variation Detection in the Human*
1037 *Genome.* Curr Protoc Hum Genet, 2020. **107**(1): p. e103.
1038 59. Ewels, P.A., et al., *The nf-core framework for community-curated bioinformatics*
1039 *pipelines.* Nat Biotechnol, 2020. **38**(3): p. 276-278.
1040 60. Patel, H., et al., *nf-core/rnaseq: nf-core/rnaseq v3.1.0*. Vol. 3.1.0. 2024.
1041 61. Hart, T., et al., *Finding the active genes in deep RNA-seq gene expression studies.*
1042 BMC Genomics, 2013. **14**: p. 778.
1043 62. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and*
1044 *dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
1045 63. Wu, T., et al., *clusterProfiler 4.0: A universal enrichment tool for interpreting omics*
1046 *data.* Innovation (Camb), 2021. **2**(3): p. 100141.
1047 64. Saul, D., et al., *A new gene set identifies senescent cells and predicts senescence-*
1048 *associated pathways across tissues.* Nat Commun, 2022. **13**(1): p. 4827.
1049 65. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.*
1050 Nucleic Acids Res, 2000. **28**(1): p. 27-30.
1051 66. Gundry, M., et al., *Direct, genome-wide assessment of DNA mutations in single*
1052 *cells.* Nucleic Acids Res, 2012. **40**(5): p. 2032-40.
1053 67. *FastQC*. 2015.
1054 68. Krueger, F., *Trim galore.* A wrapper tool around Cutadapt and FastQC to
1055 consistently apply quality and adapter trimming to FastQ files, 2015. **516**(517): p.
1056 517.
1057 69. DePristo, M.A., et al., *A framework for variation discovery and genotyping using*
1058 *next-generation DNA sequencing data.* Nature Genetics, 2011. **43**(5): p. 491-498.
1059 70. Petljak, M., et al., *Characterizing Mutational Signatures in Human Cancer Cell*
1060 *Lines Reveals Episodic APOBEC Mutagenesis.* Cell, 2019. **176**(6): p. 1282-
1061 1294.e20.
1062 71. Zhang, L., et al., *Analysis of Somatic Mutations in Senescent Cells Using Single-*
1063 *Cell Whole-Genome Sequencing.* Aging Biol, 2023. **1**(1).
1064 72. Blokzijl, F., et al., *MutationalPatterns: comprehensive genome-wide analysis of*
1065 *mutational processes.* Genome Med, 2018. **10**(1): p. 33.
1066 73. Sun, S., et al., *SomaMutDB: a database of somatic mutations in normal human*
1067 *tissues.* Nucleic Acids Res, 2022. **50**(D1): p. D1100-d1108.
1068 74. Sondka, Z., et al., *COSMIC: a curated database of somatic variants and clinical*
1069 *data for cancer.* Nucleic Acids Res, 2024. **52**(D1): p. D1210-d1217.
1070 75. Degasperi, A., et al., *A practical framework and online tool for mutational signature*
1071 *analyses show inter-tissue variation and driver dependencies.* Nat Cancer, 2020.
1072 **1**(2): p. 249-263.
1073 76. Selby, C.P., et al., *Molecular Mechanisms of Transcription-Coupled Repair.* Annu
1074 Rev Biochem, 2023. **92**: p. 115-144.
1075 77. Broad, D., *DepMap 24Q4 Public*. 2024, Figshare+.
1076 78. Blomen, V.A., et al., *Gene essentiality and synthetic lethality in haploid human*
1077 *cells.* Science, 2015. **350**(6264): p. 1092-6.
1078 79. Lawrence, M., et al., *Software for computing and annotating genomic ranges.*
1079 PLoS Comput Biol, 2013. **9**(8): p. e1003118.
1080 80. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids
1081 Res, 2001. **29**(1): p. 308-11.

1082 81. Krasovec, M., et al., *Spontaneous Mutation Rate in the Smallest Photosynthetic*
1083 *Eukaryotes.* Mol Biol Evol, 2017. **34**(7): p. 1770-1779.

1084

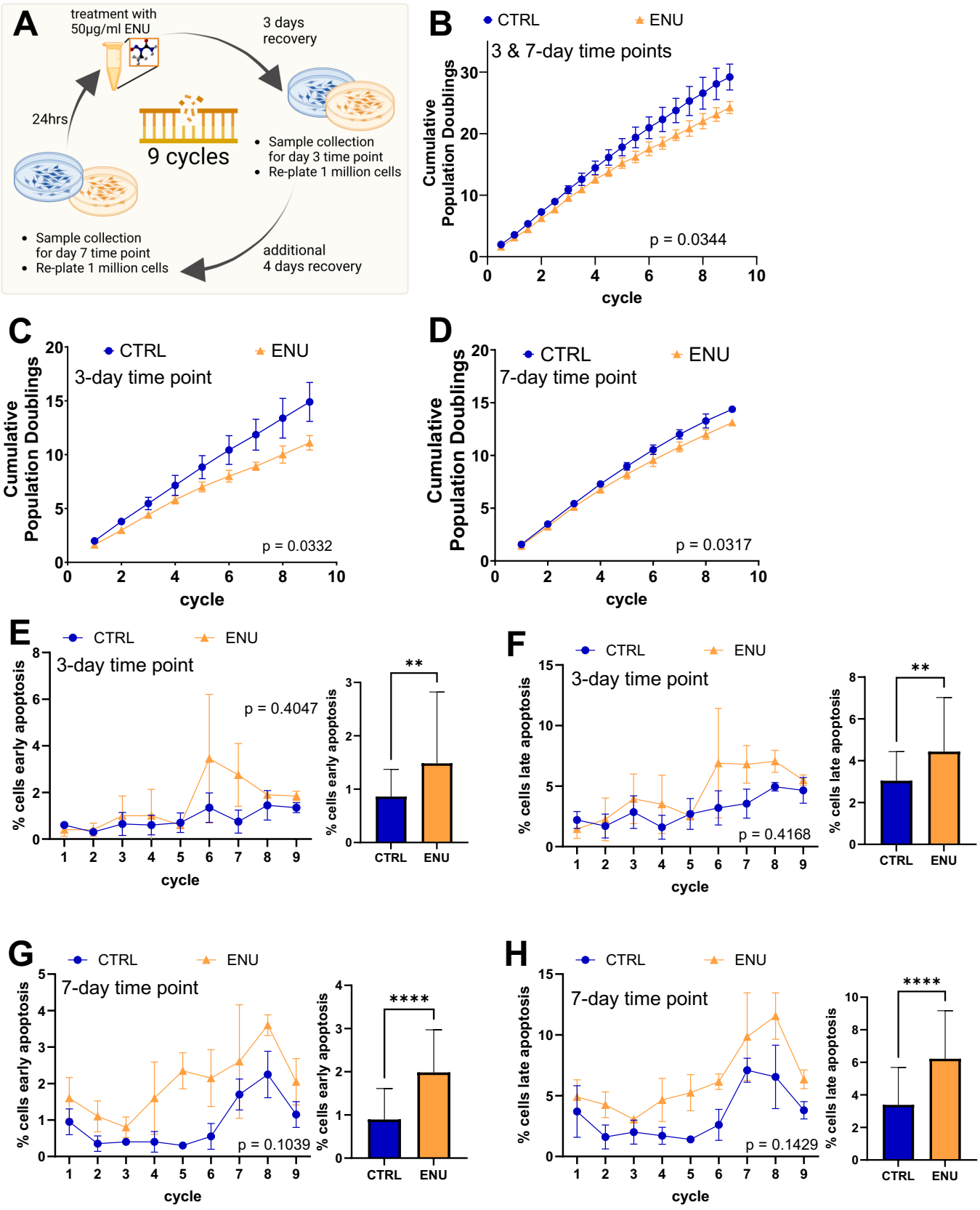


Figure 1. Repeated Mutagen Treatment of Normal Cells Only Slightly Affects Cell Growth

(A) Schematic depiction of experimental design. Human fetal lung fibroblasts (IMR-90) were treated with a sublethal dose of 50 µg/mL of N-ethyl-N-nitrosourea (ENU), which was followed by a recovery period of 7 days. At 3-day and 7-day time points, 1 million cells were passaged and a sample was taken for further analysis. This process was repeated for 9 cycles.

(B) Cumulative population doublings (PDs) of the control (CTRL) and ENU-treated (ENU) groups throughout the experiment. Statistics: n=3; Data represent mean ± S.D.; P-value estimated using a two-way repeated measures ANOVA.

(C) Cumulative PDs of the CTRL and ENU groups assessed at the 3-day time point. Statistics as in (B).

(D) Cumulative PDs of the CTRL and ENU groups at the 7-day time point. Statistics as in (B).

(E) Percentage of cells detected in early apoptosis at the 3-day time point at each cycle (left) or with all cycles pooled (right). Statistics: n=2; Data represent mean ± S.D.; P-value of data from each cycle was estimated using a two-way repeated measures ANOVA. P-value of data from all cycles pooled was estimated using a paired t-test; **: p <= 0.005; ****: p<= 0.0001.

(F) Percentage of cells detected in late apoptosis at the 3-day time point at each cycle (left) or with all cycles pooled (right). Statistics as in (E).

(G) Percentage of cells detected in early apoptosis at the 7-day time point at each cycle (left) or with all cycles pooled (right). Statistics as in (E).

(H) Percentage of cells detected in late apoptosis at the 7-day time point at each cycle (left) or with all cycles pooled (right). Statistics as in (E).

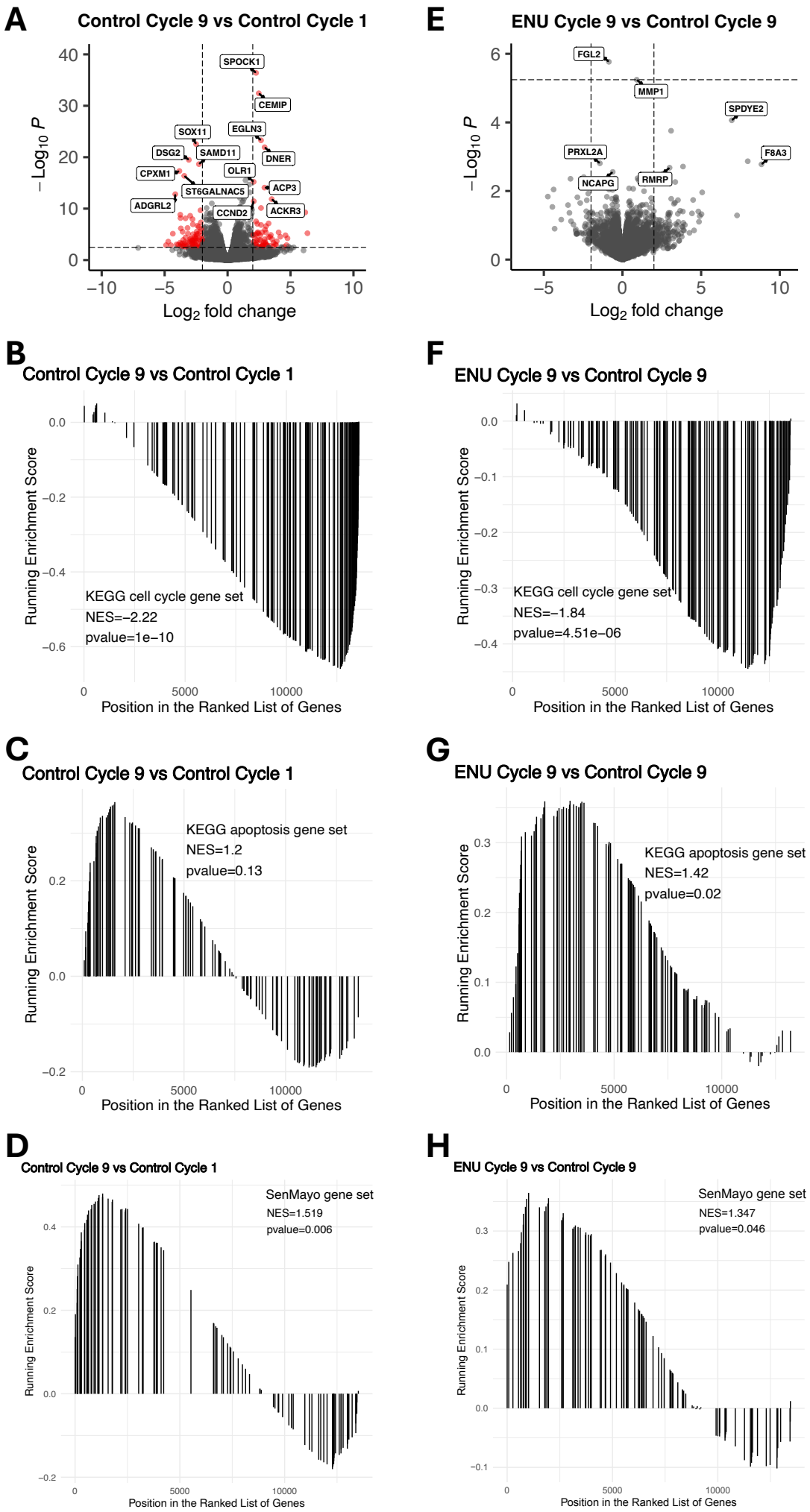


Figure 2. RNA-Sequencing Reveals Increased Cell Cycle Inhibition and Apoptosis in ENU-treated Cells

(A) Differential gene expression (DGE) of the control cycle 9 vs. control cycle 1 comparison. Red points in are Log_2 fold change>2 and adjusted p-value<0.05. See Table S2 for detailed results. Statistics: DGE results obtained using Wald test followed by Benjamini-Hochberg correction for multiple hypothesis testing.

(B-D) Gene set enrichment analysis (GSEA) of DGE results from the control cycle 9 vs. control cycle 1 comparison for KEGG cell cycle (B), KEGG apoptosis (C), and SenMayo senescence (D) gene sets. Statistics: NES = normalized enrichment score, p-values obtained using a permutation test.

(E) Same as in (A), but for the ENU-treated cycle 9 vs. control cycle 9 comparison. See Table S2 for detailed results. Statistics: Same as in (A).

(F-H) Same as in (B-D), but for the ENU-treated cycle 9 vs. control cycle 9 comparison. Statistics: Same as in (B-D).

Statistics for all panels: n=3. See Table S1 for sample details.

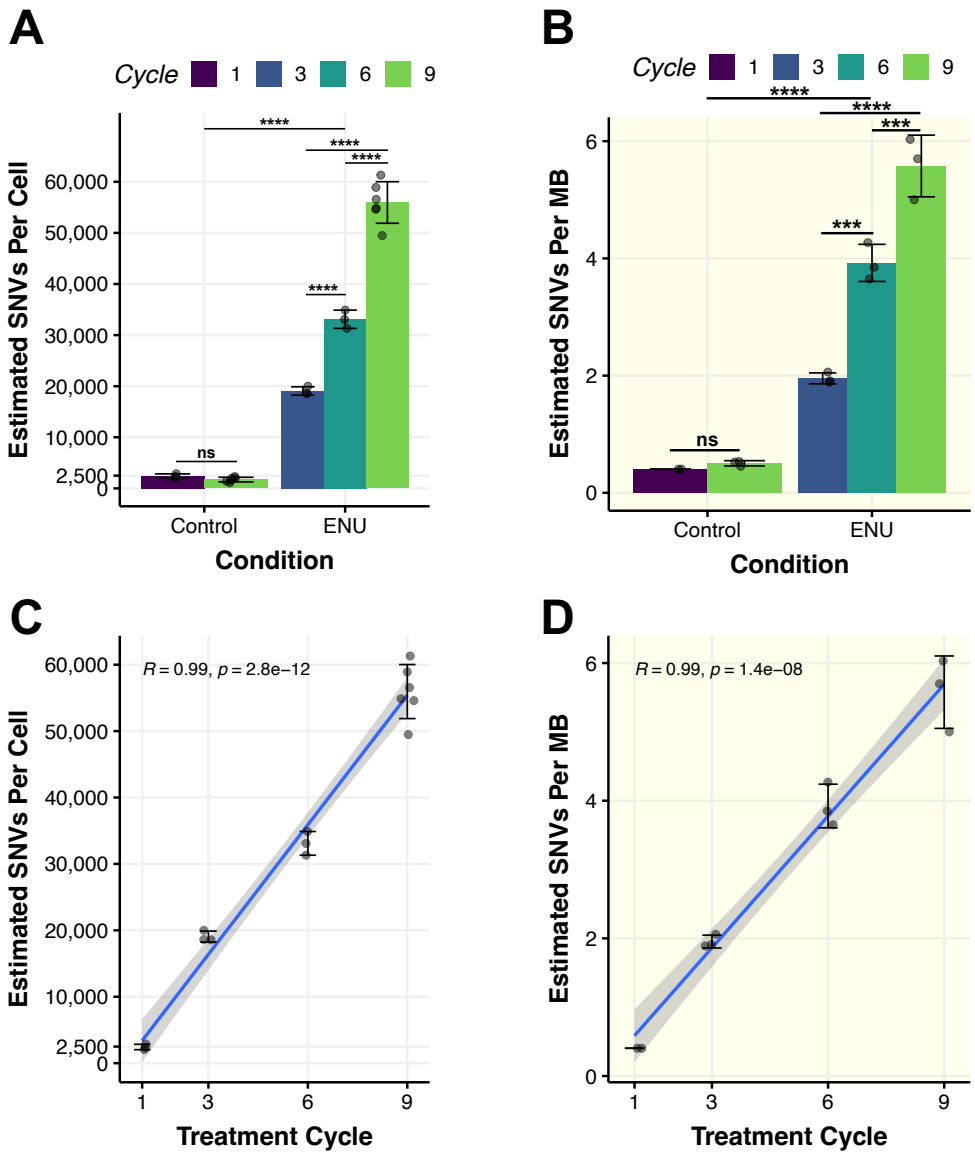


Figure 3. Repeated Mutagen Treatment of Normal Cells Results in Extremely High Mutation Burden.

(A) Estimated load of somatic single nucleotide variants (SNVs) per cell corrected for genome coverage and sensitivity. See figure S2H for observed SNV load per cell. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D. p-values of comparisons between cycles within the same condition were estimated with a two-sided linear model with estimated marginal means, Tukey's HSD. P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and shown at very top of plot. p-value legend: ****: p <= 0.0001.

(B) Estimated frequency of somatic SNVs per megabase (MB) using the SMM-Seq method. See supplemental note 2 for details of this analysis. See figure S3A for comparison of single-cell and single-molecule mutation burden estimates. Statistics: n=3; same as in (A).

(C) Fit of linear model to estimated SNV load per cell shown in (A) vs. the treatment cycle cells were collected at. Control cycle 9 is omitted. Shown at top is the fit of a linear equation. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; data represent mean ± S.D.; Pearson R correlation coefficient and p-value of the correlation was estimated using a T-test.

(D) Fit of linear model to estimated frequency of SNVs per MB shown in (B) vs. the treatment cycle cells were collected at. Control cycle 9 is omitted. Shown at the top is the fit of a linear equation. Statistics: n=3; Same as in (C).

See Tables S3-4 for sample details.

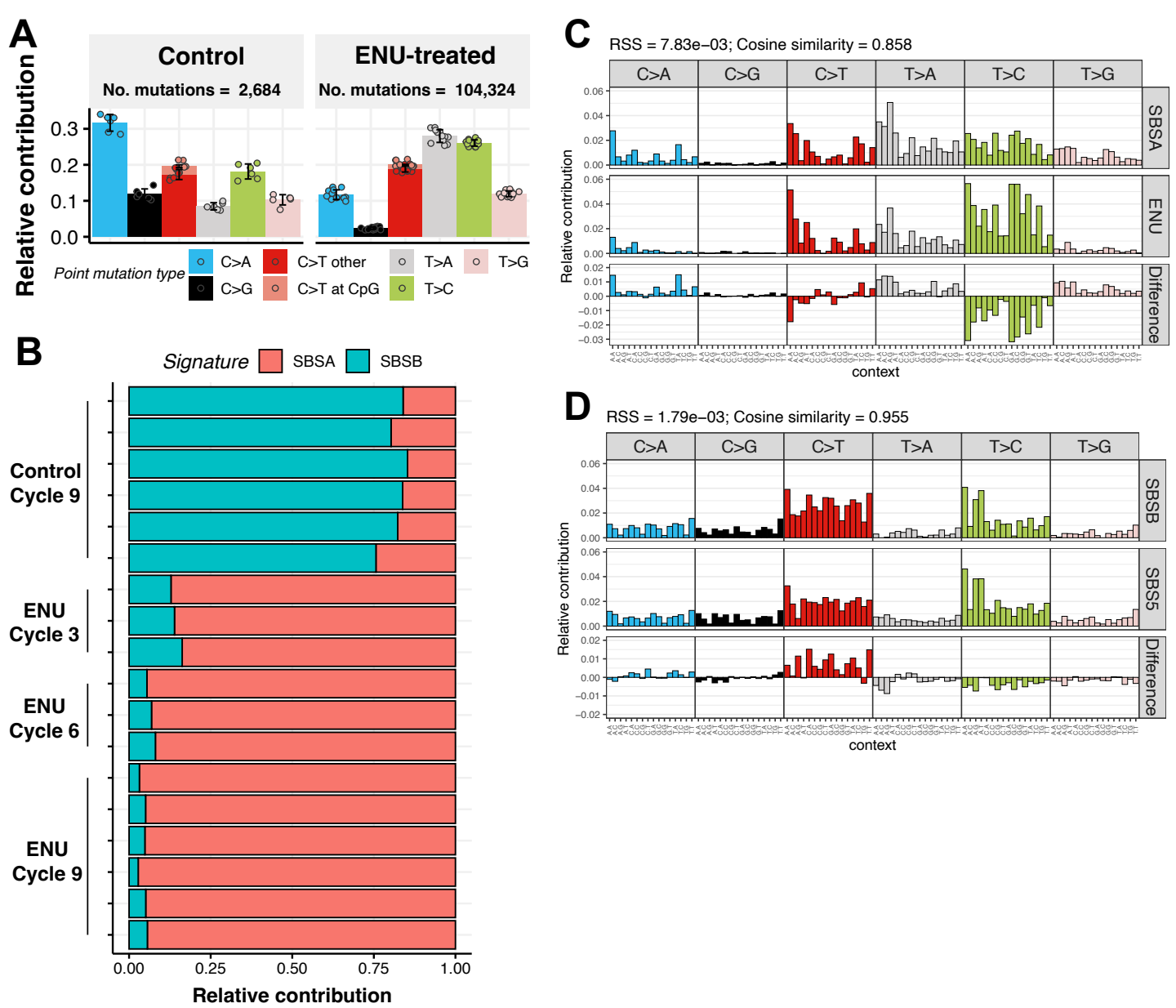


Figure 4. Mutational Spectra And Signatures Resulting From Repeated Mutagen Treatment
(A) Mutational spectra of the relative contribution of SNV types per cell grouped by condition. See figure S4A for groups stratified by cycle. Note that cells from the control cycle 1 group were excluded from the mutational spectra and signatures analyses due to abnormally high C>T frequency suggestive of a sample preparation artifact (Figures S4F-G; Methods). Statistics: n=3 cells for cycles 3 and 6; n=6 cells for cycle 9; Data represent mean ± S.D.
(**B)** Relative contribution of *de novo* mutational signatures SBSA and SBSB to individual cells (shown as rows) grouped by condition and cycle. See figure S4I for determination of number of signatures to extract.
(C) Comparison of *de novo* mutational signature SBSA to the ENU signature from the SIGNAL database. Statistics: cosine similarity 0.858, residual sum of squares=$7.83*10^{-3}$.
(D) Comparison of *de novo* mutational signature SBSB to most similar fitted signature SBS5 from the COSMIC database. Statistics: cosine similarity=0.955, residual sum of squares= $1.79*10^{-3}$.

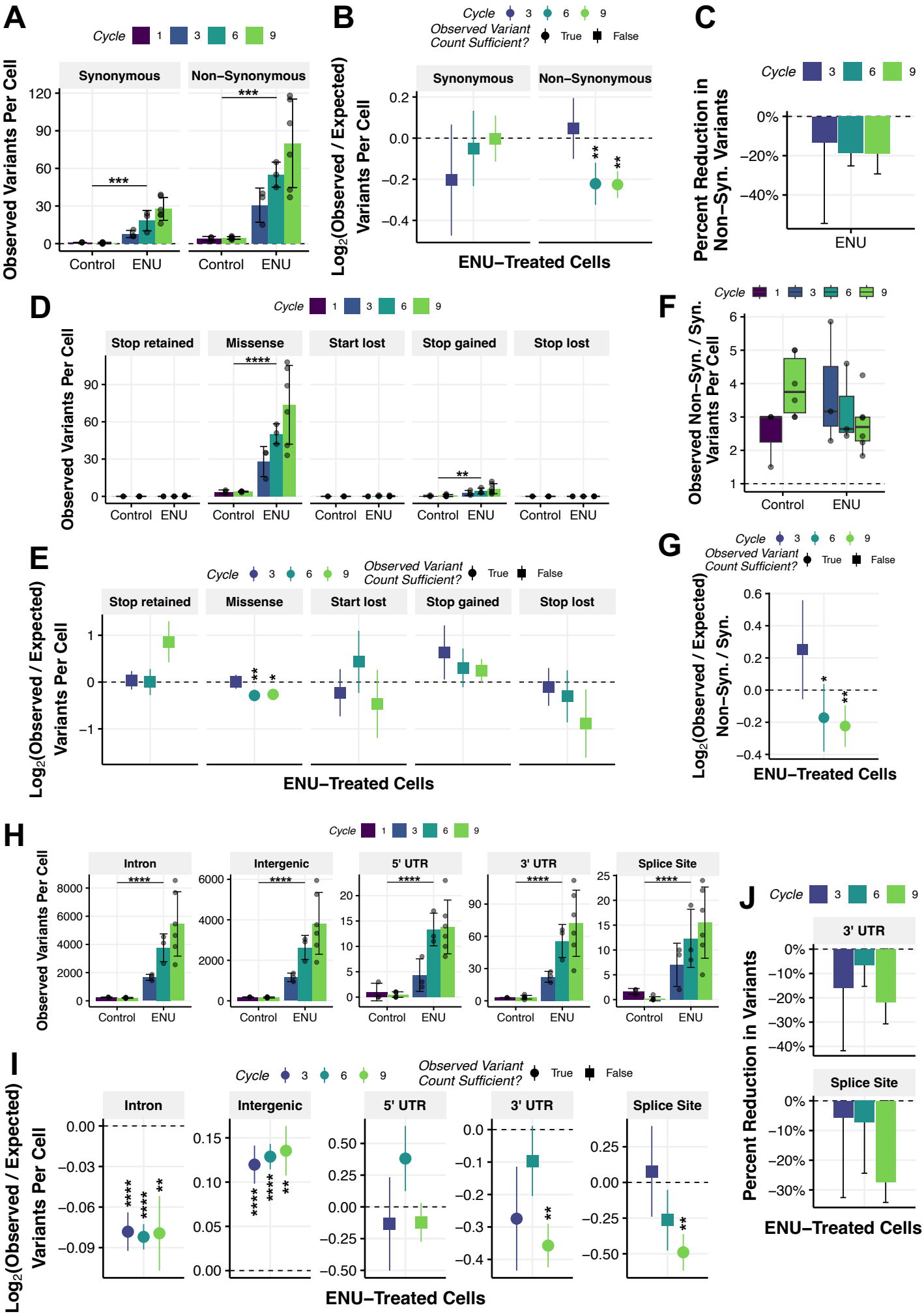


Figure 5. Negative Selection Against Damaging Coding and Non-Coding Variants in ENU-treated Cells.
(A) Observed frequency of synonymous and non-synonymous variants per cell. Statistics: P-values for comparisons of control and ENU-treated groups were estimated using a two-way ANOVA.
(B) Log_2(O/E) ratio of synonymous and non-synonymous variants for ENU-treated cells. See figures S5C, S5D, and S5E for results of control cells, without normalization, and per cell, respectively. Statistics: P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000). P-values were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S5B; Table S5).
(C) Estimated percent of non-synonymous variants out of total coding variants that are protected against in the ENU-treated group due to negative selection.
(D) Observed frequency of various types of non-synonymous variants per cell. Statistics: Same as in (A)
(E) Log_2O/E of various types of non-synonymous variants for ENU-treated cells. See figures S5F, S5G, and S5H for results of control cells, without normalization, and per cell, respectively. Statistics: Same as in (B).
(F) Ratio of observed non-synonymous to synonymous (N/S) variants per cell.
(G) Log_2O/E of the N/S ratio for ENU-treated cells. See figures S5I and S5J for results of control cells and per cell, respectively. Statistics: Same as in (B).
(H) Observed frequency of various types of non-coding variants per cell. Statistics: Same as in (A)
(I) Log_2O/E of various types of non-coding variants for ENU-treated cells. See figures S5K, S5L, and S5M for results of control cells, without normalization, and per cell, respectively. See Supplementary Note 5 for discussion on results of the intergenic region. Statistics: Same as in (B).
(J) Estimated percent reduction of variants in 3' UTR and splice region variants that are protected against in the ENU-treated group due to negative selection.
Statistics for all panels: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; Data represent mean ± S.D.; p-value legend: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001

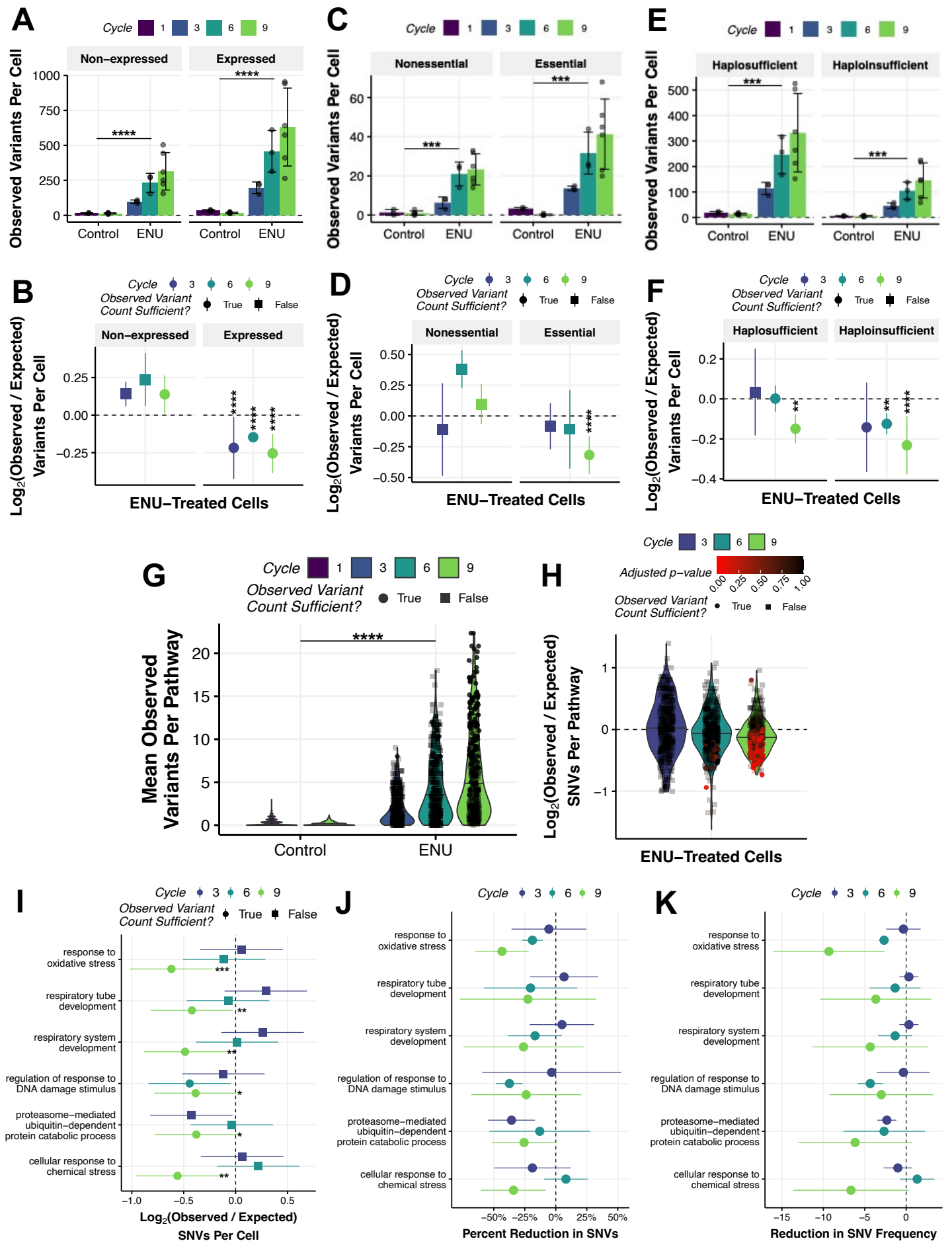


Figure 6. Pathways Supporting Fibroblast Growth and Cell Identity are Protected from Mutation Accumulation

(A) Observed frequency of SNVs per cell in functional regions associated with genes that were determined to be non-expressed or expressed from bulk mRNA transcriptome data (Table S7). Statistics: P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA.

(B) Log_2(O/E) ratio of SNVs in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data for ENU-treated cells. See figures S7A, S7B, and S7C for results of control cells, without normalization, and per cell, respectively. Statistics: P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000). P-values were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S5B; Table S9).

(C) Observed frequency of SNVs per cell in functional regions associated with expressed genes belonging to nonessential or essential gene sets for ENU-treated cells. Statistics: Same in in (A).

(D) Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to nonessential or essential gene sets for ENU-treated cells. See figures S7D, S7E, and S7F for results of control cells, without normalization, and per cell, respectively. Statistics: Same as in (B).

(E) Observed frequency of SNVs per cell in functional regions associated with expressed genes belonging to haplosufficient or haploinsufficient gene sets for control and ENU-treated cells. Statistics: Same in in (A).

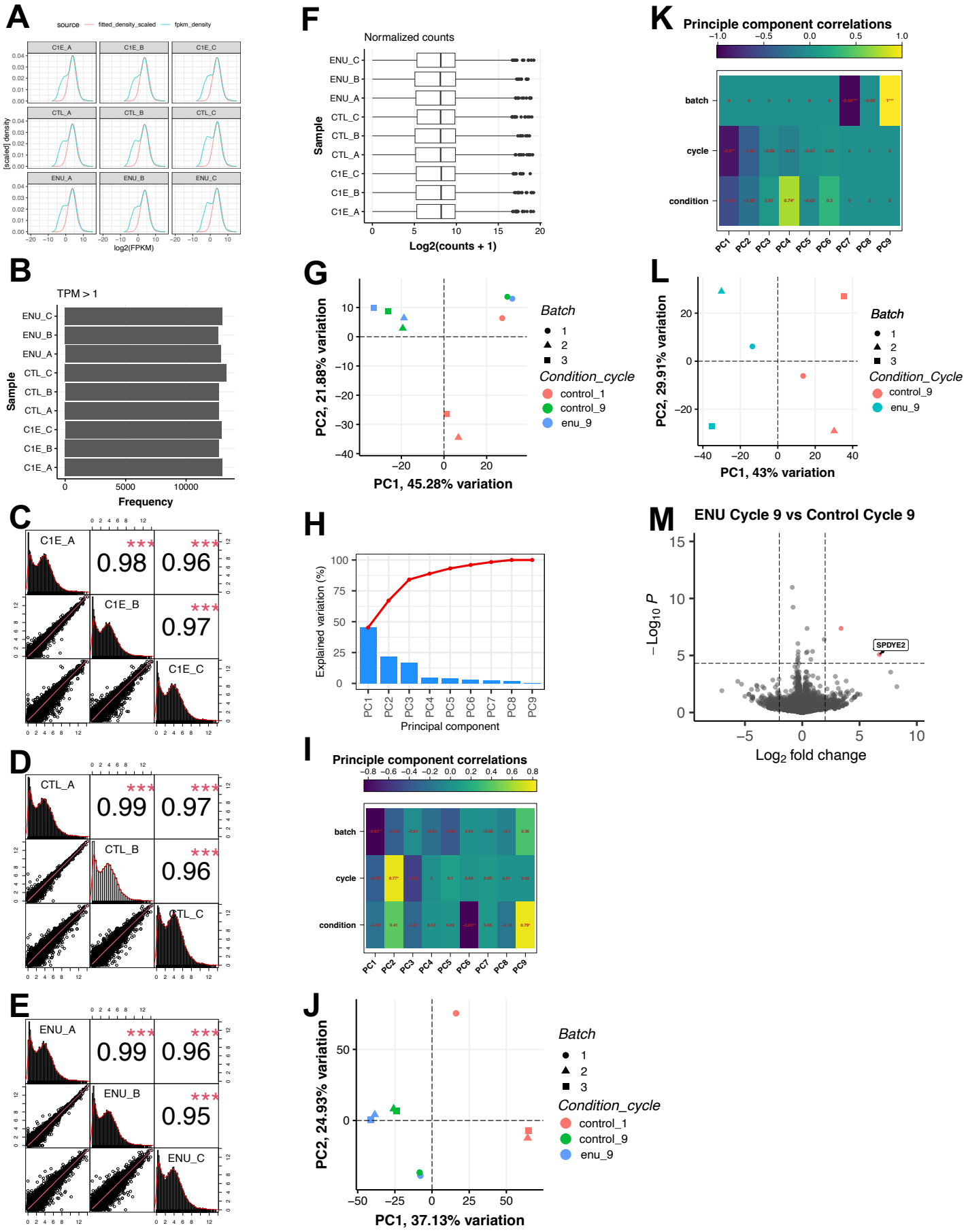
(F) Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to haplosufficient or haploinsufficient gene sets for ENU-treated cells. See figures S7G, S7H, and S7I for results of control cells, without normalization, and per cell, respectively. Statistics: Same as in (B).

(G) Mean observed frequency of SNVs within the functional regions of expressed pathways. See figure S7J for top 20 most enriched pathways. Statistics: n=399 pathways; P-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA.

(H) Mean Log_2(O/E) ratio of SNVs within the functional regions of expressed pathways. See figure S7K and S7L for results of control cell and without normalization, respectively. Statistics: n=399 pathways; Same as in (B), with the addition of Benjamini-Hochberg correction for multiple hypothesis testing of pathways.

(I-K) Mean Log_2(O/E) ratio (I), estimated percent reduction in SNVs (J), and estimated reduction in SNV frequency (K) in functional regions of pathways found to be under significant negative selection at cycle 9 in the ENU-treated group.

Statistics for all panels: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; Data represent mean ± S.D.; p-value legend: ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001



Supplementary Figure 1 | Transcriptome Analysis Quality Control and Extended Analysis. Related to Figure 2.

(A) Gene abundance distribution of each sample showing unfiltered distribution (blue) and the Gaussian fit to the Log_2(FPKM) transcript abundances, which represents the filtering threshold used to identify expressed genes used in downstream analyses.

(B) Number of genes detected at TPM > 1.

(C-E) Correlation of Log_2(TPM) data between biological replicates in control cycle 1 (C), control cycle 9 (D), and ENU-treated cycle 9 (E) groups. Statistics: n=16,133; Pearson R correlation coefficient; p-value were estimated using a t-test; p-value legend: ***: p <= 0.001.

(F) Distribution of DESeq2 normalized Log2(counts+1) transcript abundances.

(G) Principle component analysis (PCA) of all samples.

(H) Percent of variance explained by each PC.

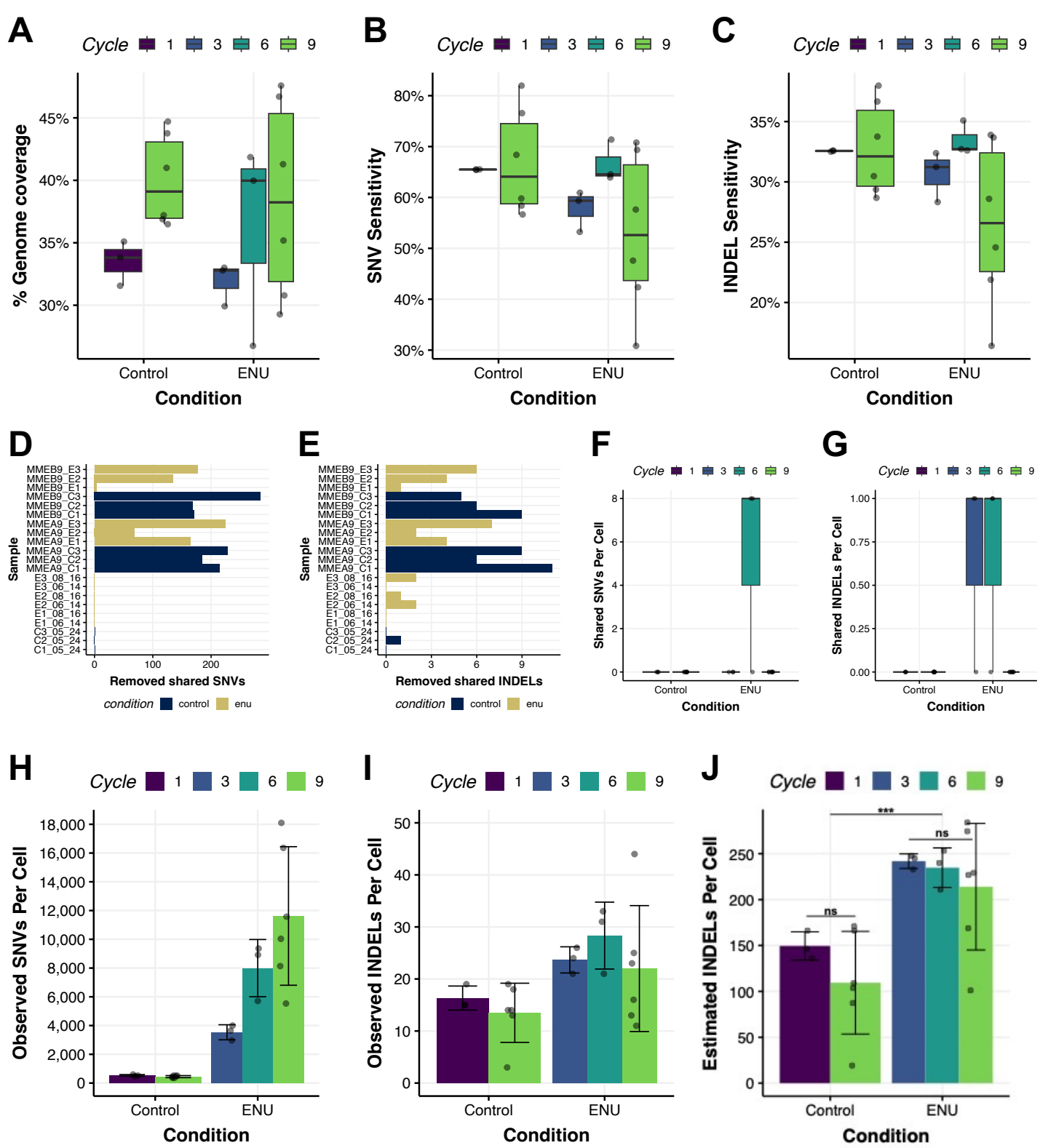
(I) Correlation of each PC to experimental variables, showing the batch effect that is captured by the first PC. Statistics: n=3; Pearson R correlation coefficient; p-value were estimated using a t-test. **: p <= 0.01.

(J) PCA of all samples after adjusting for the batch variable (Methods).

(K) Correlation of each PC to experimental variables after batch correction, showing a lack of correlation of the batch variable with the first 6 PCs. Statistics: Same as in (I).

(L) PCA of samples only from cycle 9 after adjusting for the batch variable (Methods). See supplemental note 1 for details on the purpose of this analysis.

(M) DGE results of ENU-treated cycle 9 vs. control cycle 9 when only analyzing cycle 9 samples after adjusting for the batch variable. Red points are Log_2 fold change>2 and adjusted p-value<0.05. Statistics: n=3; DGE results obtained using Wald test followed by Benjamini-Hochberg correction for multiple hypothesis testing.



Supplementary Figure 2 | Single-Cell Whole Genome Analysis Quality Control and Observed Mutational Burden. Related to Figure 3.

(A) Percentage of genome covered at 20X read depth per cell.

(B-C) Sensitivity of SNV (B) and INDEL (C) mutation calling (Methods).

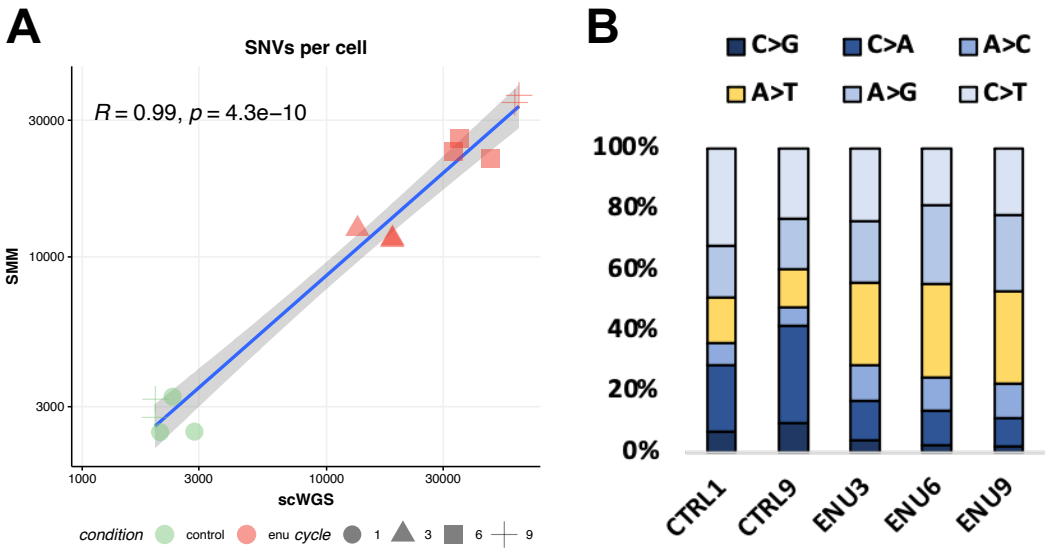
(D-E) Number of observed SNVs (D) and INDELs (E) removed from cells collected in the same cycle and batch that shared SNVs and INDELs (Methods).

(F-G) Number of observed shared SNVs and INDELs amongst all cells after filtering for mutations shared within each cycle and batch (Methods).

(H-I) Observed SNVs (H), INDELs (I) per cell. Statistics: Data represent mean ± S.D

(J) Estimated burden of somatic short insertions/deletions and (INDELs) per cell corrected for genome coverage and sensitivity. Statistics: Data represent mean ± S.D; p-values of comparisons between cycles within the same condition were estimated with a two-sided linear model with estimated marginal means, Tukey's HSD. p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA and are shown at top of plots; p-value legend ***: p <= 0.001

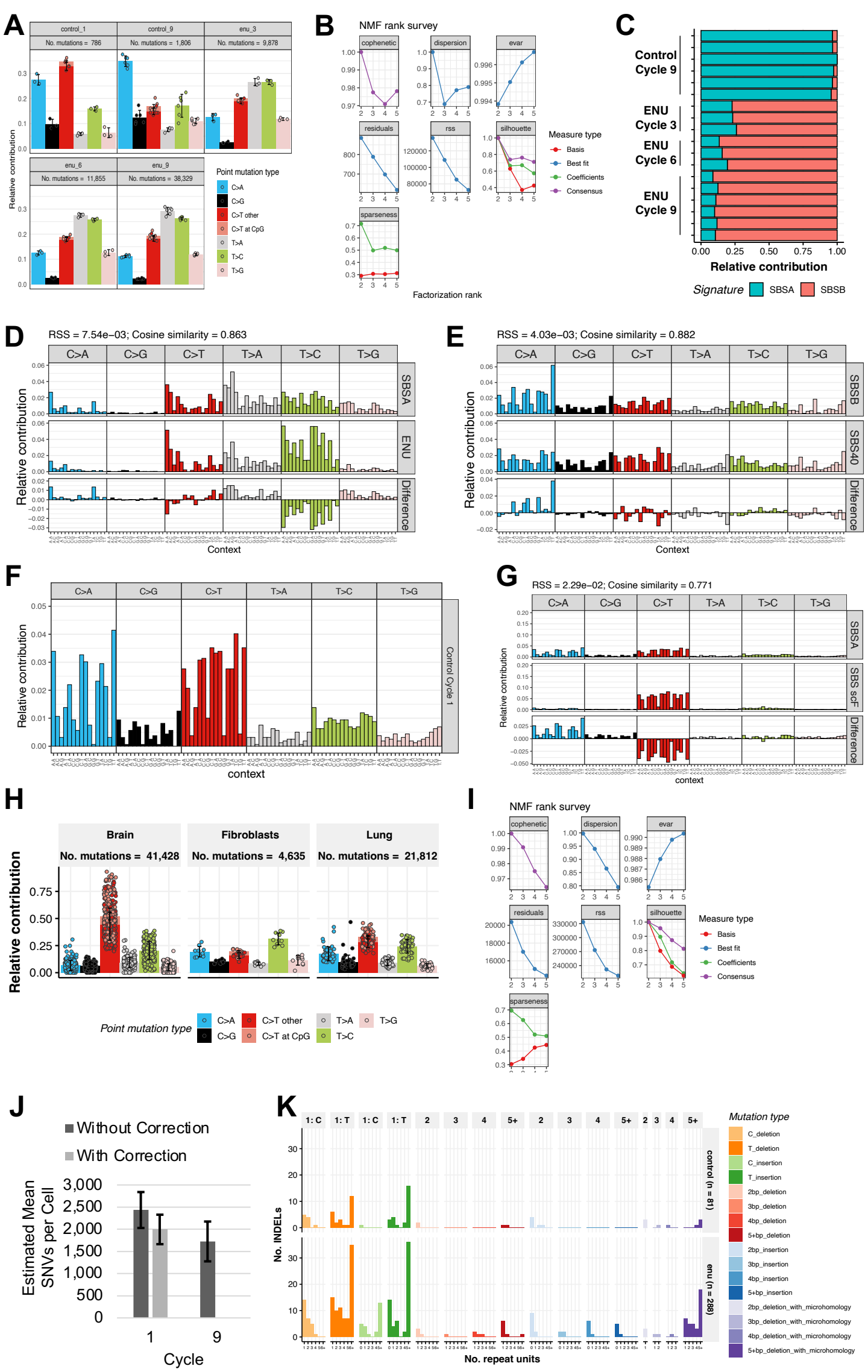
Statistics for all panels: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9



Supplementary Figure 3 | Comparison of Single-Cell and Single-Molecule Whole Genome Analyses. Related to Figure 3.

(A) Comparison of estimated SNV burden between the single-cell whole genome sequencing (scWGS) and single-molecule mutation sequencing (SMM-seq) assays matched by biological replicate. X- and Y-axis shown on Log_{10} scale. Related to figures 3A-B. Statistics: n=2-3; Shown at top is Pearson R correlation coefficient and p-value of the correlation that was estimate using a T-test.

(B) Mutation spectra of the relative contribution of SNV types for all samples within the same cycle and condition. Note the increase in the proportion of T>A base transversions, which is characteristic of ENU-induced mutations. Related to figure 3B. Statistics: n=2-3



Supplementary Figure 4 | Extended analysis of Mutational Spectra and Signatures. Related to Figure 4.

(A) Mutation spectra of the relative contribution of SNV types per cell separated by cycle and condition. Note the abnormal increase of C>T transition indicative of a technical artifact in control cycle 1 group, which was removed from downstream signature analysis (Methods). Related to figure 4A. Statistics: Data represent mean ± S.D.; n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9

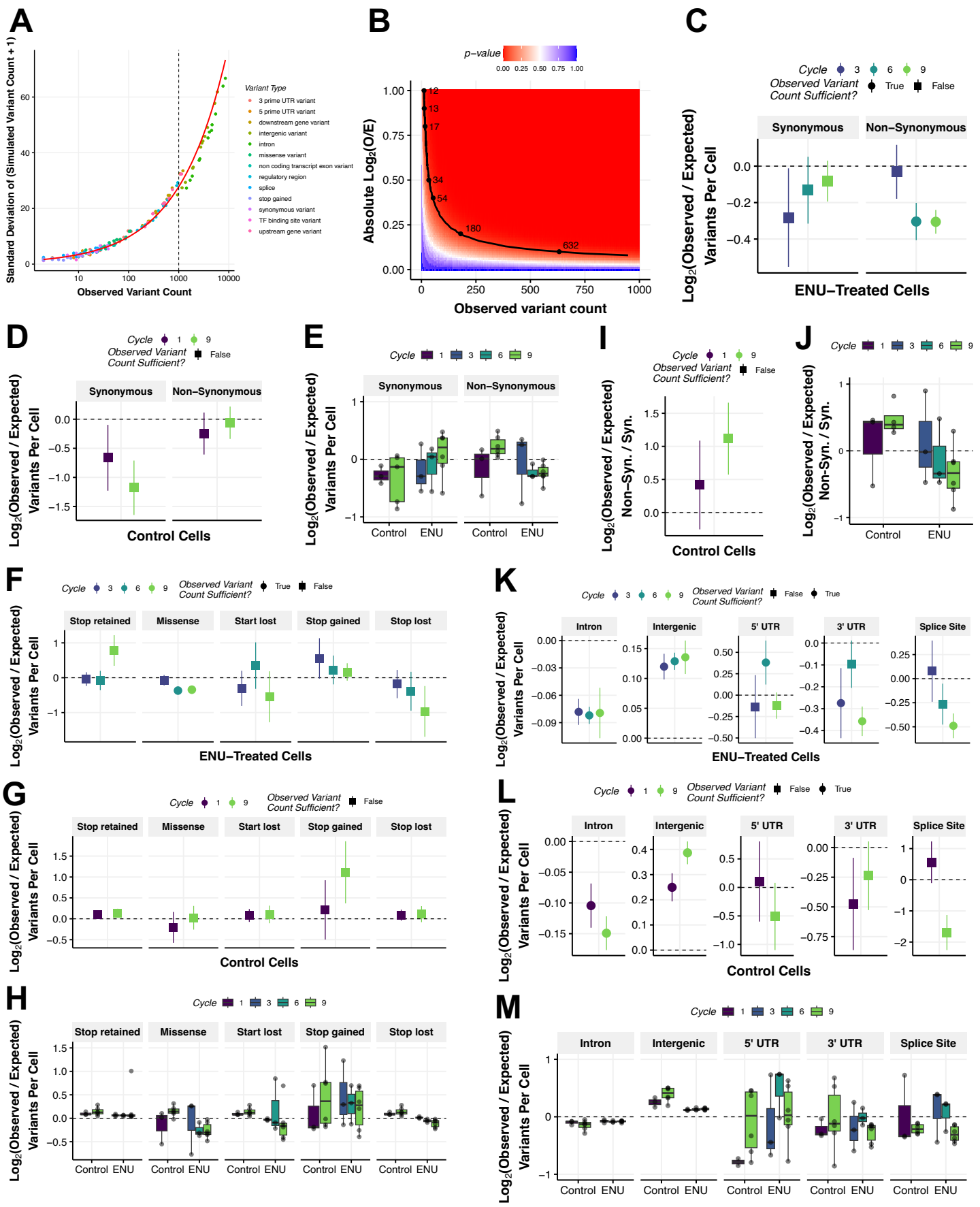
(B) Results of determining optimal factorization rank for SNV signature extraction using non-negative matrix factorization, where a rank of 2 was chosen to extract 2 mutational signatures from the mutation data. Related to figure 4B.

(C) Mutational profile of control cycle 1 group, showing higher than expected cytosine-to-thymine (C>T) transversions at GCN contexts. Statistics: n=3

(D) Comparison of the only mutational signature extracted from the control cycle 1 group, SBSA, to a known single cell library preparation artifact SBS scF. See Methods for details on the exclusion of this group from this analysis. Statistics: n=3; cosine similarity=0.771, residual sum of squares=$2.29*10^{-2}$

(E) Estimated burden of somatic single nucleotide variants (SNVs) per cell in the control groups corrected for genome coverage, sensitivity, and the scF sequencing artifact mutational signature (C-D). This artifact increased the proportion of C>T transitions by 18% ± 1%. With this, we estimated the amount of artifactual C>T transitions in control cells at cycle 1, subtracted this from the observed mutation burden, and then re-estimated the SNVs per cell. Statistics: n=3 cells for cycle 1; n=6 cells for cycle 9; data represent mean ± S.D.

(F) INDEL mutation spectra grouped by condition. See supplemental note 3 for detailed description of each spectrum. Statistics: n=6 cells for control group; n=12 cells for ENU-treated group.



Supplementary Figure 5 | Extended Analysis of Negative Selection Against Damaging Coding and Non-Coding Variants in ENU-treated Cells. Related to Figure 5.

(A) Standard deviation of expected (simulated) variants as a function of observed variant frequency. The data was fit to a gamma GLM with a log link (red line) Statistics: Residual deviance = 3.19 on 144 df vs. null deviance = 147.37 on 145 df; dispersion = 0.020; log-transformed observed variant count was a highly significant predictor of expected variant standard deviation, irrespective of the variant type ($p < 2 \times 10^{-16}$).

(B) P-value as a function of simulations of the observed variant count and the Log_2(O/E) to determine the number of observed variants given the Log_2(O/E) that is needed to reliably interpret the resulting statistical test.

(C) Unnormalized Log_2(O/E) ratio of synonymous and non-synonymous variants for ENU-treated cells, related to figure 5B.

(D) Normalized Log_2(O/E) ratio of synonymous and non-synonymous variants for control cells, related to figure 5B.

(E) Normalized Log_2(O/E) ratio of synonymous and non-synonymous variants per cell, related to figure 5B.

(F) Unnormalized Log_2(O/E) ratio of coding variants for ENU-treated cells, related to figure 5E.

(G) Normalized Log_2(O/E) ratio of coding variants for control cells, related to figure 5E.

(H) Normalized Log_2(O/E) ratio of coding variants per cell, related to figure 5E.

(I) Normalized Log_2(O/E) ratio of the non-synonymous / synonymous (N/S) ratio for control cells, related to figure 5G.

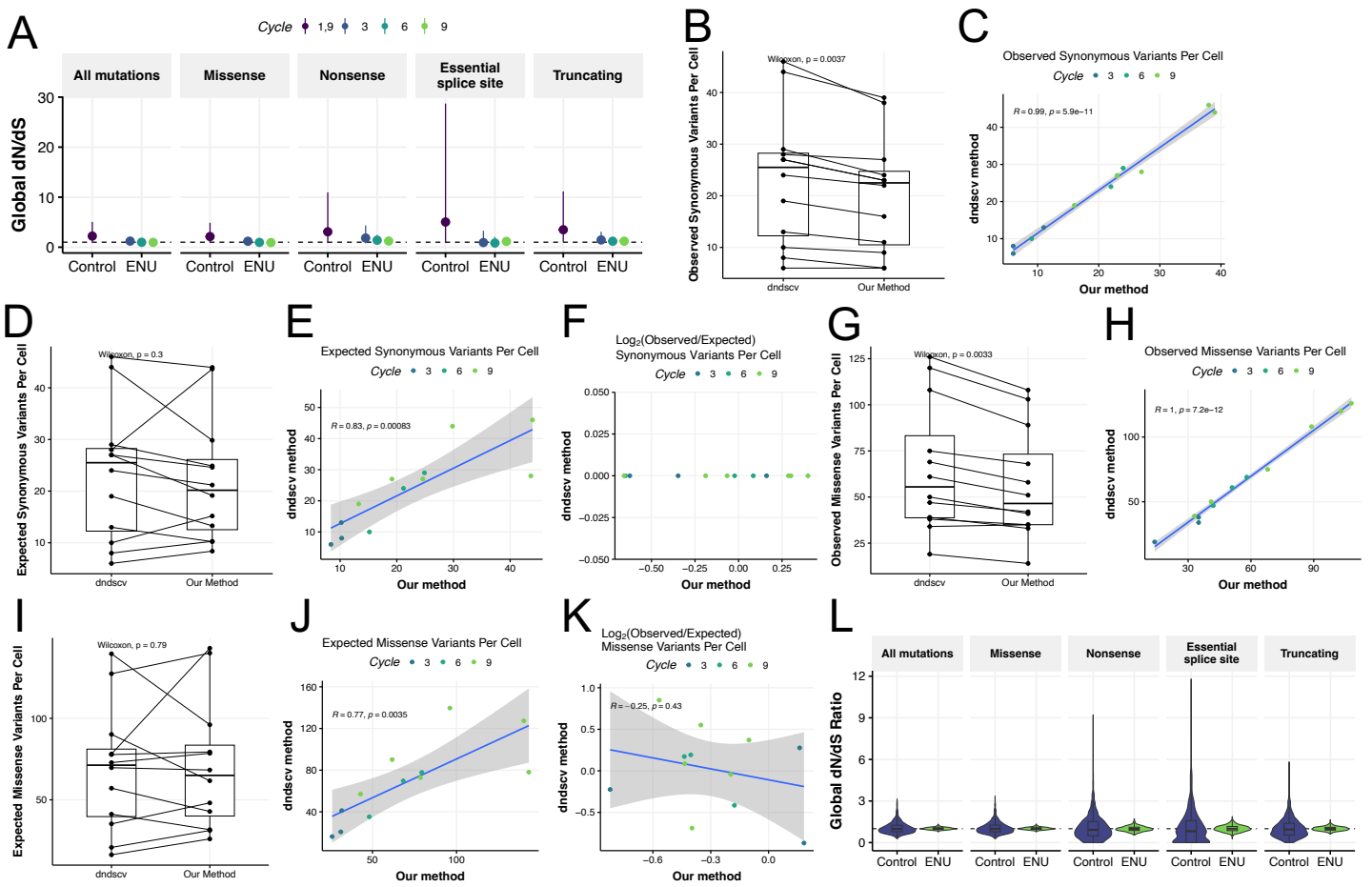
(J) Normalized Log_2(O/E) ratio of the non-synonymous / synonymous (N/S) ratio per cell for control and ENU-treated cells, related to figure 5G.

(K) Unnormalized Log_2(O/E) ratio of non-coding variants for ENU-treated cells, related to figure 5I.

(L) Normalized Log_2(O/E) ratio of non-coding variants for control cells, related to figure 5I.

(M) Normalized Log_2(O/E) ratio of non-coding variants per cell, related to figure 5I.

Statistics for all panels: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; Data represent mean ± S.D.



Supplementary Figure 6 | dN/dS analysis of selection pressure for damaging coding variants. Related to Figure 5.

(A) Standard deviation of expected (simulated) variants as a function of observed variant frequency. The data was fit to a gamma GLM with a log link (red line) Statistics: Residual deviance = 3.19 on 144 df vs. null deviance = 147.37 on 145 df; dispersion = 0.020; log-transformed observed variant count was a highly significant predictor of expected variant standard deviation, irrespective of the variant type ($p < 2 \times 10^{-16}$).

(B) P-value as a function of simulations of the observed variant count and the Log_2(O/E) to determine the number of observed variants given the Log_2(O/E) that is needed to reliably interpret the resulting statistical test.

(C) Unnormalized Log_2(O/E) ratio of synonymous and non-synonymous variants for ENU-treated cells, related to figure 5B.

(D) Normalized Log_2(O/E) ratio of synonymous and non-synonymous variants for control cells, related to figure 5B.

(E) Normalized Log_2(O/E) ratio of synonymous and non-synonymous variants per cell, related to figure 5B.

(F) Log_2(O/E) ratio of the non-synonymous / synonymous (N/S) ratio for control cells, related to figure 5G.

(G) Same as in (F), but per cell, related to figure 5G.

(H) Unnormalized Log_2(O/E) ratio of coding variants for ENU-treated cells, related to figure 5E.

(I) Normalized Log_2(O/E) ratio of coding variants for control cells, related to figure 5E.

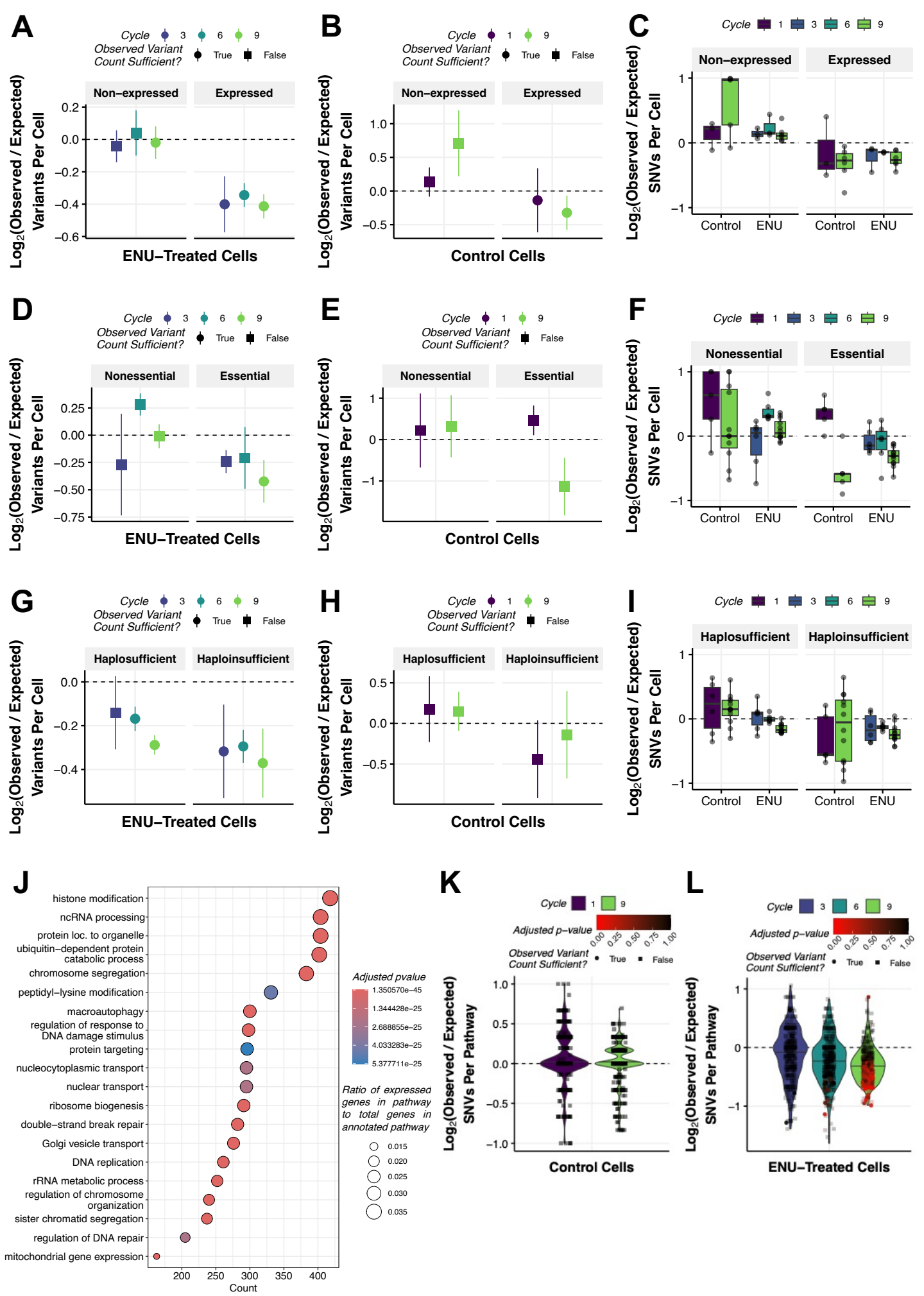
(J) Normalized Log_2(O/E) ratio of coding variants per cell, related to figure 5E.

(K) Unnormalized Log_2(O/E) ratio of non-coding variants for ENU-treated cells, related to figure 5I.

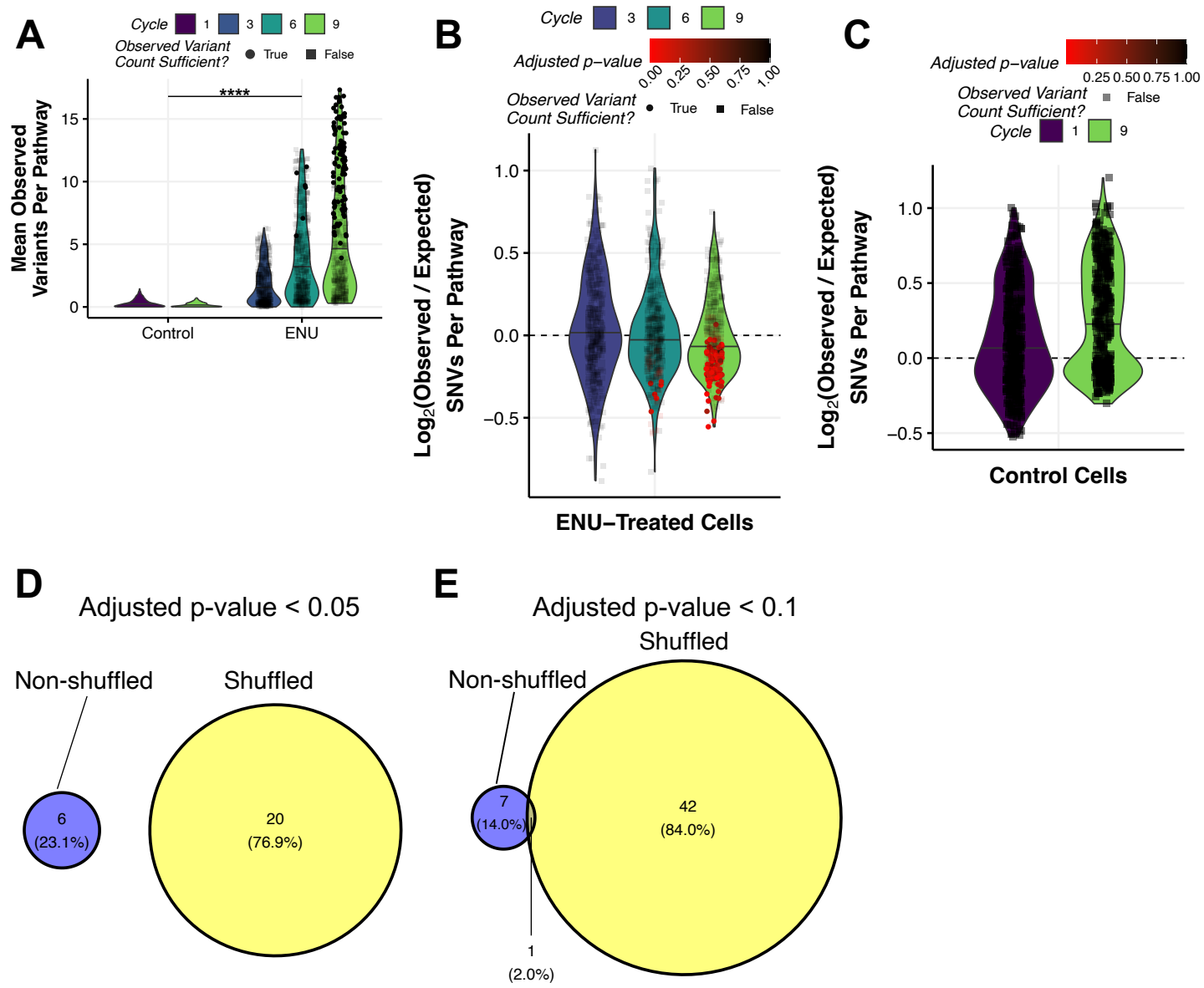
(L) Normalized Log_2(O/E) ratio of non-coding variants for control cells, related to figure 5I.

(M) Normalized Log_2(O/E) ratio of non-coding variants per cell, related to figure 5I.

Statistics for all panels: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; Data represent mean ± S.D.



Supplementary Figure 7 | Extended Selection Pressure Analysis of Mutation Accumulation at the Pathway Level. Related to Figure 6.
(A) Unnormalized Log_2(O/E) ratio of SNVs in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data for ENU-treated cells. Related to figure 6B. Statistics: n=3 cells for cycles 1, 3, and 6; n=6 cells for cycle 9; Data represent mean ± S.D.
(B) Normalized Log_2(O/E) ratio of SNVs in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data for control cells. Related to figure 6B. Statistics same as in (A).
(C) Normalized Log_2(O/E) ratio of SNVs per cell in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data for control and ENU-treated cells. Related to figure 6B. Related to figure 6B. Statistics same as in (A).
(D) Unnormalized Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to nonessential or essential gene sets for ENU-treated cells. Related to figure 6C. Related to figure 6B. Statistics same as in (A).
(E) Normalized Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to nonessential or essential gene sets for control cells. Related to figure 6C. Related to figure 6B. Statistics same as in (A).
(F) Normalized Log_2(O/E) ratio of SNVs per cell in functional regions associated with expressed genes belonging to nonessential or essential gene sets for control and ENU-treated cells. Related to figure 6C. Related to figure 6B. Statistics same as in (A).
(G) Unnormalized Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to haplosufficient or haploinsufficient gene sets for ENU-treated cells. Related to figure 6F. Related to figure 6B. Statistics same as in (A).
(H) Normalized Log_2(O/E) ratio of SNVs in functional regions associated with expressed genes belonging to haplosufficient or haploinsufficient gene sets for control cells. Related to figure 6F. Related to figure 6B. Statistics same as in (A).
(I) Normalized Log_2(O/E) ratio of SNVs per cell in functional regions associated with expressed genes belonging to haplosufficient or haploinsufficient gene sets for control and ENU-treated cells. Related to figure 6F. Related to figure 6B. Statistics same as in (A).
(J) Top 20 gene ontology biological processes out of 399 resulting from gene set over representation analysis of expressed genes from bulk mRNA transcriptome data from control and ENU-treated cells at cycle 9. Statistics: n=6; adjusted p-values estimated using a hypergeometric distribution test followed by Benjamini-Hochberg correction for multiple hypothesis testing. Log_2 O/E ratio of SNVs per cell in functional regions associated with genes determined to be non-expressed or expressed from bulk mRNA transcriptome data from control and ENU-treated cells at cycle 9. Related to figures 6G-H.
(K) Unnormalized Log_2(O/E) ratio of SNVs in the functional regions of expressed pathways for ENU-treated cells. Related to figure 6G. Statistics: P-values for comparing the frequency of observed variants to that expected by chance alone were estimated using a permutation test (n=10,000) with Benjamini-Hochberg correction for multiple hypothesis testing.
(L) Normalized Log_2(O/E) ratio of SNVs in the functional regions of expressed pathways for control cells. Related to figure 6G. Statistics: Same as in (K).



Supplementary Figure 8 | Negative Control to Validate Negative Selection of Functional Pathways. Related to Figure 6.

(A) Mean observed frequency of SNVs within the functional regions of expressed pathways after shuffling genes between pathways (negative control). Statistics: n=399 pathways; p-values of comparisons of control and ENU-treated groups were estimated using a two-way ANOVA.

(B) Mean Log_2(O/E) ratio of SNVs within the functional regions of expressed pathways for ENU-treated cells after shuffling genes between pathways (negative control). Statistics: P-values to assess if the Log_2(O/E) ratio significantly deviates from 0 were estimated using a permutation test (n=10,000) with the addition of Benjamini-Hochberg correction for multiple hypothesis testing of pathways. P-values were only estimated if the frequency of observed variants allowed for sufficiency statistical power (Figure S5B).

(C) Normalized Log_2(O/E) ratio of SNVs in the functional regions of expressed for control cells. pathways after shuffling genes between pathways (negative control). Statistics: Same as in (B).

(D-E) Overlap of pathways that show significant negative O/E ratios (i.e. pathways under negative selection) at cycle 9 in the ENU-treated group at an adjusted p-value<0.05 threshold (C) or an adjusted p-value<0.1 threshold (D) resulting from the original non-shuffled analysis or shuffled analysis (negative control).